The Evolutionary Consequences of Phenotypic Mutations

Dissertation

zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Siniša Bratulić
aus
Kroatien

Promotionskomitee
Prof. Dr. Andreas Wagner (Vorsitz)
Prof. Dr. Lukas Keller
Prof. Dr. Martin Ackermann

Zürich, 2016

Contents

Summary

Translation is a key cellular process during which ribosomes use mRNA as a template for protein synthesis. Yet, translation is remarkably error-prone. Mistranslation occurs when ribosomes misread mRNA and incorporate incorrect amino acids into the nascent protein. These errors, called phenotypic mutations, can destabilize proteins and promote protein misfolding. Misfolded proteins tend to form protein aggregates. Protein aggregates are cytotoxic, and they are causative factors in human neurodegenerative diseases. Selection against misfolding is thought to be a major factor influencing the evolution of protein stability.

Surprisingly, mistranslation can also be beneficial in certain conditions. Mistranslation increases proteomic diversity, which can help populations that are faced with an environmental challenge. For example such diversity can help pathogens evade the immune response during an infection or survive treatments with certain antibiotics. Because phenotypic mutations have physiological, ecological, and evolutionary consequences, they have attracted a lot of interest in molecular evolution and cell biology. In this thesis I studied mistranslation by pursuing three projects.

First, I studied the evolutionary adaptation of proteins to mistranslation. Specifically, I wanted to see whether proteins adapt to mistranslation by adopting synonymous changes that locally increase translational accuracy, or by adopting nonsynonymous changes that increase protein stability. To this end, I experimentally evolved the antibiotic resistance gene TEM-1 in *Escherichia coli* hosts with either normal or elevated rates of mistranslation. I found that under selection with low concentrations of antibiotic, mistranslating populations mitigate mistranslation-induced costs by reducing protein expression. Under selection with high concentrations of antibiotics, mistranslation led to accumulation of nonsynonymous substitutions that stabilize TEM-1.

In the second project, I studied how mistranslation affects the evolution of resistance to a new antibiotic. Specifically, I evolved TEM-1 in *E. coli* hosts with either normal or high rates of mistranslation, with selection for cefotaxime. Using a similar experimental design as in the first project, I showed that mistranslation increases the genetic diversity of bacterial populations. Furthermore, I found that this genetic diversity can help bacterial populations adapt to antibiotics other than cefotaxime.

In the third part, I focused on characterizing mistranslation rates. I used preexisting mass spectrometry proteomics datasets from two pathogenic bacterial species to directly identify mistranslated proteins. I showed that mistranslation can create radical amino acid changes at high frequencies, and that these changes affect many essential proteins. More importantly, some of the highly mistranslated proteins are needed for virulence and pathogenesis, and were identically mistranslated in both bacterial species. This suggests that phenotypic mutations might benefit pathogenic bacteria.

In sum, my research was aimed at addressing some long-standing questions in molecular evolution. My findings suggest that mistranslation could be more common than previously thought, and that it can influence the evolution of antibiotic resistance and proteomic diversity in bacteria.

Zusammenfassung

Translation ist ein wesentlicher zellulärer Prozess, in dem Ribosome die mRNA als Vorlage für Proteinsynthese benutzen. Dennoch ist die Translation bemerkenswert fehleranfällig. Zu einer fehlerhaften Translation kommt es, wenn Ribosome die mRNA falsch lesen und dann falsche Aminosäuren in ein naszierendes Protein einbauen. Diese Fehler, die Phänotyp-Mutationen genannt werden, können Proteine destabilisieren und Proteinfehlfaltung ("misfolding") fördern. Üblicherweise bilden fehlgefaltete Proteine Proteinaggregate. Proteinaggregate sind zytotoxisch und ursächliche Faktoren neurodegenerativer Erkrankungen bei Menschen. Selektion gegen Fehlfaltung gilt als einer der wichtigsten Faktoren, die die Evolution der Proteinstabilität beeinflusst.

Überraschenderweise kann fehlerhafte Translation unter bestimmten Umständen auch von Vorteil sein. Fehlerhafte Translation erhöht proteomische Vielfalt, was für Populationen, die Umweltherausforderungen gegenüberstehen, hilfreich sein kann. Beispielsweise kann solche Vielfalt Krankheitserregern dabei helfen, der Reaktion des Immunsystems bei einer Infektion auszuweichen, oder die Behandlung mit bestimmten Antibiotika zu überleben. Da phänotypische Mutationen physiologische, ökologische und evolutionäre Folgen haben, sind sie von Interesse in der Molekularevolution und Zellbiologie. In dieser Dissertation habe ich die fehlerhafte Translation in drei Projekten untersucht.

Zunächst habe ich die evolutionäre Anpassung der Proteine an fehlerhafte Translation untersucht. Insbesondere wollte ich feststellen, ob sich Proteine an fehlerhafter Translation durch synonyme Veränderungen, die die Translationsgenauigkeit lokal erhöhen, anpassen, oder durch Übernahme nicht-synonymer Veränderungen, die die Proteinstabilität erhöhen. Zu diesem Zweck habe ich das antibiotikaresistente Gen TEM-1 in *Escherichia coli* Wirten mit entweder normalen oder erhöhten fehlerhaften Translationsraten experimentell evolviert. Ich habe festgestellt, dass bei Selektion mit einer niedrigen Antibiotika-konzentration, Populationen mit fehlerhafter Translation die durch Translationfehler verursachten Kosten durch reduzierte Proteinexpression vermindert werden. Bei Selektion mit hoher Antibiotika-konzentration, verursacht fehlerhafte Translation die Anhäufung von nicht-synonymen Substitutionen, die TEM-1 stabilisieren.

Im zweiten Projekt untersuchte ich wie fehlerhafter Translation die Evolution der Resistenz gegen neue Antibiotika beeinflusst. Insbesondere habe ich TEM-1 in *E. coli* Wirten mit normalen und erhöhten Fehltranslationsraten unter Selektion auf Zefotaxim evolviert. Mittels eines ähnlichen Experiments wie im ersten Projekt, habe ich gezeigt, dass fehlerhafter Translation die genetische Vielfalt bakterieller Populationen steigert. Darüber hinaus fand ich, dass diese genetische Vielfalt der bakteriellen Populationen bei der Anpassung an weitere Antibiotika helfen kann.

Im dritten Projekt habe ich mich auf die Charakterisierung der Fehltranslationsraten konzentriert. Ich habe mich der vorher vorhandenen Datensätze der massenspektrometrischen Proteomik aus zwei pathogenen bakteriellen Spezies bedient, um fehlerhaft translatierte Proteine direkt zu identifizieren. Ich habe gezeigt, dass fehlerhafte Translation häufig radikale Aminosäureveränderungen verursachen kann, und dass diese Veränderungen viele wesentliche Proteine beeinflussen können. Noch wichtiger ist es, dass einige der fehltranslatierten Proteine für Virulenz und Pathogenese nötig sind, und in beiden bakteriellen Gattungen identisch

fehltranslantiert wurden. Dies legt nahe, dass phänotypische Mutationen für pathogene Bakterien von Vorteil sein können.

Zusammenfassend lässt sich feststellen, dass meine Forschungsergebnisse nahe legen, dass Fehltranslation viel häufiger als bisher angenommen sein könnte, und dass sie die Evolution der Antibiotikaresistenz und der proteomischen Diversität in Bakterien beeinflussen kann.

Chapter 1

Introduction

1.1 Molecular noise in biological systems

Living systems, and their ability to adapt to the environment, often evoke the impression of optimal design and perfection. However, biophysical and energetic constraints impose limits on how close to "perfection" molecular systems can exist [1]. Molecular events that regulate cellular processes often involve interactions between few molecules, which causes stochastic effects. Moreover, the discrimination between correct and incorrect molecular interactions can be unreliable because it is often based on weak forces [1]. Consequently, biomolecular processes suffer from limited accuracy and heterogeneity. The resulting random variability of quantities such as protein concentrations within a population of genetically identical individuals is called *cellular* or *molecular noise*. Noise leads to heterogeneities even in isogenic populations and homogeneous environments [2].

The idea that biochemical systems are affected by stochasticity and molecular noise was predicted from basic physical principles already in 1944 by Scrödinger [3]. The biological significance of stochasticity was only demonstrated in subsequent theoretical studies, which suggested that stochastic effects in gene expression [4] could be important for creating cellular heterogeneity [2], and for regulating development [5]. However, experimental studies directly linking molecular stochasticity to physiological effects became possible only a little over a decade ago. Advances in experimental methods made it possible to quantify heterogeneities on a single-cell level in tightly controlled environments [6, 7].

1.2 Physiological and evolutionary consequences of noise

How molecular noise affects phenotypes has been the focus of intense study in the last decade [8–11]. Because phenotypic heterogeneity affects reproduction and survivial, molecular noise can directly affect the evolutionary fate of a genotype. Molecular noise is often viewed as being detrimental to cellular fitness [10, 11]. However, noise can also be beneficial, since it increases phenotypic diversity [12], and drives cellular differentiation and development [8, 9].

Noise and heterogeneity in gene expression are by far the most well studied examples of biological noise [13, 14], but noise affects all cellular processes [1]. For example, consider a single protein-coding gene. In a population of organisms, this gene might be polymorphic due to genetic mutations. In addition, the amino-acid composition of the expressed protein may even be heterogeneous in a single cell, and differ from the sequence encoded by the gene because of errors in transcription [15] and translation [1, 16]. Moreover, the concentration of the encoded mRNA and protein might differ among isogenic cells because of the stochastic nature of gene expression [6]. Such varying protein concentrations might cause variations in metabolite concentrations and growth rates [17], as well as differences in developmental programs [9, 14].

Most noise is probably detrimental and natural selection may thus act to reduce it [10].

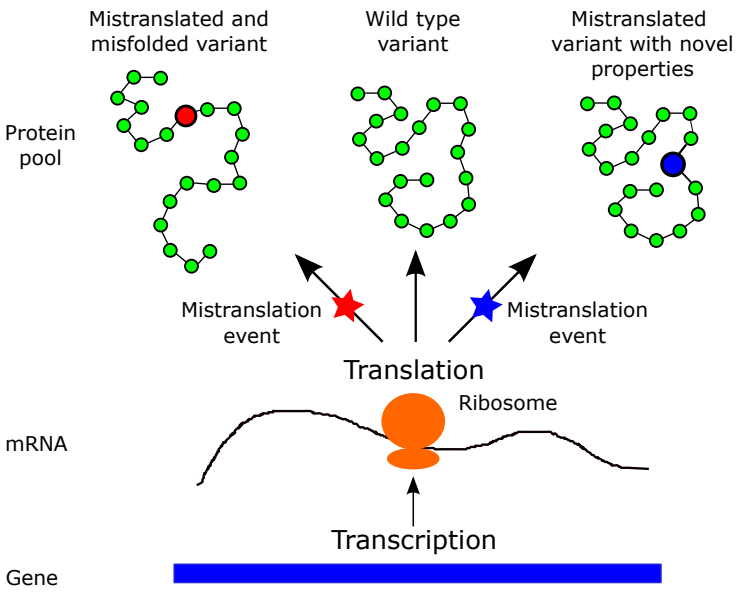

Figure 1.1: Mistranslation and the statistical protein pools. Mistranslation events create a pool of diverse protein products from a single gene because they introduce phenotypic mutations into proteins. On the one hand, phenotypic mutations can be deleterious (red circle) because they can destabilize proteins and cause misfolding. On the other hand, phenotypic mutations can impart novel biochemical and biophysical properties to mistranslated proteins (blue circle).

However, there are examples where noise is beneficial and can help survival. For example, in *Saccharomyces cerevisiae*, noisy gene expression can help cells survive high levels of antibiotic Zeocin [18]. Moreover, noise can be domesticated and used in cellular decision making. Cellular decision making is a process where cells assume different heritable phenotypes, for example during development [14]. Gene regulatory networks that control phenotypic switching are often bistable and sensitive to changes in mRNA or protein concentrations [19]. This enables cells to exploit noisy expression to create subpopulations of phenotypically diverse cells in a population of genetically identical cells [14]. This noise-induced phenotypic switching might provide a large fitness benefit in changing environments, because one of the subpopulations might find itself better adapted when the environment changes [20, 21].

In this thesis, I will focus on a particular type of noise, namely *phenotypic mutations*. I will describe how such mutations arise, their physiological consequences, and how they affect the evolution of proteins and cellular systems.

1.3 Phenotypic mutations and their rates

Phenotypic mutations are transient changes that occur when genes (DNA) are expressed, i.e. transcribed and translated into mRNA and protein, respectively. Translation is thought to be the most error-prone part of gene expression [22] because ribosomes can mistranslate mRNA, causing missense, nonsense, frameshift, and readthrough errors at high rates [16, 22]. One consequence of phenotypic mutations is that they create a pool of *statistical proteins* [23] that subtly differ in their sequence, structure, and function (figure 1.1). Studying the character and the rate of phenotypic mutations is necessary to understand their physiological and evolutionary consequences [22].

Bacterial genomic mutation rates have been the subject of many studies, and typically lie between 10^{-7} and 10^{-11} per base pair per generation [24]. In contrast to genomic mutation rates, phenotypic mutation rates are orders of magnitude higher, but they are also more difficult to measure. First attempts to measure error rates in protein synthesis have been made more than 50 years ago [25]. Since then, different approaches have yielded estimates of average

phenotypic mutation rates between 10^{-4} and 10^{-3} per codon [22, 26–28]. However, we still lack a complete characterization of phenotypic error rates, including those of mistranscription, tRNA mischarging, and mistranslation across all codons even for a single species.

In the following sections I will briefly describe the different stages of gene expression, their estimated error rates, and how their fidelity is controlled. Unless specified otherwise, I will describe biochemical processes as they occur in prokaryotes.

1.3.1 Physiological consequences of phenotypic mutations

Mistranslation has real biological effects because statistical proteins can reduce the fitness of an organism [22]. In fact, some antibiotics kill bacterial cells by severely reducing the translational fidelity of their ribosomes [29]. The molecular details of how mistranslation causes deleterious effects are well described [22]. For example, mistranslated proteins often have reduced native activity, and are more sensitive to oxidative damage [30]. In addition, mistranslated proteins are destabilized and more likely to misfold [31]. Destabilized and misfolded proteins expose their hydrophobic residues and anneal them with other hydrophobic surfaces, such as other destabilized proteins or membranes. This causes the accumulation of protein-protein and protein-membrane aggregates. Such aggregates are highly toxic because they compromise the integrity of cellular membranes [22]. Furthermore, misfolded proteins can reduce viability by sequestering proteins with essential functions [32]. Finally, the high cost of synthesis and degradation of defective proteins reduces cellular fitness even further [22].

Generally, eukaryotes are less tolerant to phenotypic errors than prokaryotes [33]. In eukaryotes, increased mistranslation is linked to altered cellular morphologies [33], cell death, neurodegeneration [34], and other disease phenotypes [22, 33].

1.3.1.1 Beneficial phenotypic mutations

Several lines of evidence suggest that elevated phenotypic mutation rates can be beneficial in certain conditions. For example, the human pathogen *Candida albicans* ambiguously decodes CUG as either serine or leucine [35]. Genes coding for extracellular proteins of *C. albicans* are enriched with CUG codons. Ambiguous translation of these codons gives rise to proteins with diverse biophysical properties, modulating surface adhesion and other biologically important properties [35]. Moreover, if ambiguous decoding of CUG is induced in *Saccharomyces cerevisiae*, it triggers a general stress response, which can be adaptive in short term under conditions of elevated environmental stress [36].

In mycobacteria, mistranslation of glutamate for glutamine and aspartate for asparagine can produce phenotypic resistance to rifampicin [37]. That is because the statistical proteome produced as a consequence of these substitutions will contain a small fraction of RNA polymerase variants that are resistant to rifampicin, and thus enable cells to survive antibiotic selection.

In *Acinetobacter baylyi* mischarging (see sections 1.3.2.3) of $tRNA^{Ile}$ with valine increases the growth rate in conditions where isoleucine is limiting [38]. Also, it was recently found that *Mycoplasma* parasites have editing-defective aminoacyl-tRNA synthetases (see 1.3.3.3), causing proteome-wide mistranslation [39]. The authors hypothesized that this generates phenotypic heterogeneity in antigens, enabling parasites to evade host defense systems. Mammalian cells can also benefit from tRNA mischarging. Specifically, in the face of oxidative stress, mammalian cells can upregulate charging of non-cognate tRNAs with methionine [40]. Because methionine residues can protect proteins against oxidative damage, methionine-misacylation can be an effective strategy to protect cells against oxidative stress [40]. This response has been called *adaptive translation* [41], and has been found in all domains of life [42, 43].

Certain genes have evolved to depend on phenotypic mutations for their expression. Viruses, bacteria and yeast use programmed ribosomal frameshifting to regulate the production of different proteins encoded by a single coding region [44]. For example, in *E. coli*

the correct synthesis of DNA polymerase subunits τ and γ depends on ribosomal frameshifting [45].

Stop codon readthrough can reveal cryptic genetic diversity, and allow an organism to survive stressful conditions. Examples include the yeast prion [PSI+], which causes erroneous termination of translation, and uncovers phenotypic effects of cryptic genetic variation, which can promote survival under stress [46].

1.3.2 Gene expression

1.3.2.1 Transcription

The first step in gene expression is transcription. The transcription of information from DNA to mRNA is catalyzed by the enzyme RNA polymerase. During transcription RNA polymerase uses free ribonucleotide triphosphates as a substrate, and DNA as a template, to synthesize a mRNA molecule complementary to the template. RNA polymerase is a molecular complex consisting of four different subunits, β, β', α, and σ. The core polymerase ($\alpha_2\beta\beta'$) is responsible for catalyzing the formation of RNA, while the σ subunit serves as a factor that recognizes the transcription initiation sites known as *promoters*. Once RNA polymerase binds to a promoter, transcription starts by opening the DNA, exposing the template strand, which is then transcribed, and closed again. Transcription stops at specific regions of mRNA called transcription terminators.

1.3.2.2 Translation

Translation is a fundamental biochemical process during which the nucleotide sequence of an mRNA serves as a template to synthesize a protein with a defined amino acid sequence. The genetic code specifies the correspondence between sequences of nucleic acids and sequences of amino acids in protein products. Each of the twenty amino acids is specified by a triplet of nucleotides, called a codon. There are 64 (4^3) codons in total, 61 sense codons specifying amino acids, and 3 nonsense or stop codons specifying termination signals.

Translation is carried out by ribosomes, complex molecular machines made of three ribosomal RNAs and more than 50 proteins. These ribosomal proteins and the RNA form two ribosomal subunits, a large (50S in prokaryotes, 60S in eukaryotes), and a small (30S in prokaryotes, 40S in eukaryotes) subunit. The large subunit contains the peptidyl transferase site, which is responsible for the formation of peptide bonds in protein synthesis. The small subunit is responsible for decoding of nucleotides in mRNAs to amino acids in the nascent polypeptides during translation.

The ribosome has three sites where aminoacyl-tRNAs (aa-tRNAs, see 1.3.2.3) interact with mRNA. The A site (aminoacyl site) is used to present the aa-tRNA to the mRNA. The P site (peptidyl site) holds the nascent peptide chain bound as a peptidyl-tRNA. The E site (exit site) contains empty tRNAs that are about the exit the ribosome.

The entire process of translation involves interactions of multiple factors, and can be divided into three discrete steps.

1. *Initiation.* Initiation begins with the assembly of an initiation complex comprising a small ribosomal subunit, a mRNA, an N-formylmethyonine-tRNA, and initiation factors. The assembly of the initiation complex is directed by the so-called Shine-Dalgarno sequence upstream of the initiation codon on the mRNA. The most commonly used initiation codon is AUG, but some genes use alternative initiation codons, such as GTG. The initiation codon is paired with the initiator N-formylmethyonine-tRNA. The initiation stage is complete once the large ribosomal subunit associates with the complex, and once initiation factors are released.

2. *Elongation.* During translation, mRNA is threaded through the ribosome. The aa-tRNA is delivered to the ribosome as a part of a ternary complex with the elongation factor Tu (EF-Tu) and GTP. Upon matching of the codon to the anticodon of an aa-tRNA, EF-Tu hydrolyzes GTP into GDP, and aa-tRNA is accepted into the A-site. The movement of an aa-tRNA from

an A-site to a P-site occurs when a new peptide bond is formed. The tRNA that was bound to the nascent peptide in the previous elongation step is moved to the E-site, and the new peptidyl-tRNA takes its place. This transfer is followed by a translocation step, where the entire ribosome shifts to the next mRNA codon, and starts a new round of elongation.

3. *Termination.* The termination of protein synthesis occurs when a ribosome reaches a codon that specifies a stop signal. In the standard genetic code, this is signaled by UAG ("amber"), UAA ("ochre"), or UGA ("opal" or "umber") codons. Stop codons are recognized by release factors which cleave the polypeptide from the terminal tRNA to which it is attached. Once translation is terminated, the ribosome dissociates from mRNA, and subunits can be used to form a new initiation complex on a different mRNA template.

1.3.2.3 tRNA aminoacylation (charging)

Along with the ribosome, tRNAs are key molecules in translation. tRNAs and their cognate amino acids are linked by enzymes called aminoacyl-tRNA synthetases (aaRS). The specificity of aaRSs ensures that amino acids are paired with the correct tRNAs (anticodons) according to the genetic code.

Aminoacylation or tRNA charging is a reaction that happens in two steps. In the first step, aaRS activates an amino acid using ATP. The product of this step is an aaRS-aminoacyl-adenylate complex and inorganic pyrophosphate. In the second step, the activated aminoacyl moiety is transferred to the acceptor end of a tRNA, yielding an aminoacylated tRNA (aa-tRNA). Aminoacyl-tRNA forms a ternary complex with EF-Tu and GTP, and serves as the substrate for the elongation step in translation (see section 1.3.2.2).

1.3.3 The fidelity of gene expression

1.3.3.1 The fidelity of transcription

Selection against errors in transcription might have driven the evolution of the accuracy and specificity of RNA polymerases. The two major mechanisms that contribute to the fidelity of RNA polymerase are substrate selection [47] and proofreading of mismatched nucleotides [48]. Substrate selection depends on an element of the polymerase called the *trigger loop* [47]. The trigger loop detects the topology of RNA-DNA hybrid base pairs, and excludes mismatched ribonucleotides from the active site, preventing nucleotide misincorporation. If the incorrect substrate escapes exclusion and becomes misincorporated, transcription will be paused. This enables proofreading, i.e. the backtracking of RNA polymerase and excision of the incorrect ribonucleotide [48].

1.3.3.2 The fidelity of translation

Protein synthesis is the most expensive cellular process, consuming 30-50% of the total energy budget of the cell [49, 50]. It is therefore not surprising that there are many adaptations ensuring that protein synthesis is accurate, efficient, and tightly regulated. The fidelity of translation is controlled in three stages. The first two quality control steps occur before the peptide bond is formed, while the third step occurs thereafter. The fidelity of translation of each codon is determined by the efficiency of these quality control steps [51], and the relative concentrations of cognate aa-tRNAs in the entire pool of charged tRNAs [16].

1. *Proofreading prior to peptide bond formation.* The first quality control step is the initial selection of incoming aa-tRNA·EF-Tu·GTP ternary complex. In this step, the ternary complex is structurally distorted to allow the tRNA to efficiently scan the codon to be translated [52]. The codon-anticodon base pair matching occurs through Watson-Crick rules in the first two codon position. Matching at the third position follows rules that allow 'wobble' matches [53]. Incorrectly matched ternary complexes are preferentially discarded from the ribosome in this step. A complementary match between the codon and the anticodon causes structural

rearrangements in the decoding centre of the ribosome [54]. The structural rearrangement of the decoding centre activates the GTPase activity of the EF-Tu, leading to GTP hydrolysis [55].

Discrimination between the correct and incorrect tRNA is achieved by *kinetic proofreading* [56, 57], which is based on the difference in reaction rates between cognate and noncognate aa-tRNAs [51]. First, the dissociation rate for a noncognate ternary complex is higher than that of the cognate aa-tRNA. Even a single mismatch in the codon-anticodon complex can increase dissociation rates by 1000-fold [58]. Second, the rate of GTPase activation is lower for noncognate ternary complexes, increasing the time-widow during which they can dissociate.

Upon GTP hydrolysis, EF-Tu·GDP is released, and the aa-tRNA is accommodated into the peptidyl transferase center of the ribosome. Here, the aa-tRNA can participate in the formation of a peptide bond, or it can be rejected from the ribosome by proofreading. Again, discrimination is achieved by kinetic proofreading [51]. That is, the rate of aa-tRNA accomodation is higher for the correct aa-tRNA, and the rate of aa-tRNA rejection is higher for the incorrect aa-tRNA.

2. *Proofreading following the peptide bond formation.* Once the aa-tRNA has been accommodated in the peptidyl trasferase center, a peptide bond can be formed, and the tRNA is moved from the A- into the P-site. This is where an additional proofreading step takes place. The ribosome can recognize errors in codon-anticodon matching by monitoring the codon-anticodon helix in the P-site [59]. Mismatches in the P-site helix can increase the promiscuity of the A-site, compromising the fidelity of initial aa-tRNA selection for the next elongation step. This decreased fidelity allows release factors to promiscuously enter the A-site and to prematurely terminate the synthesis [59].

1.3.3.3 The fidelity of tRNA charging

A key quality control checkpoint in protein synthesis is the accurate pairing of an amino acid and a tRNA by an aaRS. For example, each aaRS in *E. coli* must select cognate tRNAs from as many as 86 different tRNA species [60]. In spite of this diversity, aaRSs display a high degree of substrate discrimination when compared to other enzymes [61].

As in transcription and translation, the fidelity of tRNA charging occurs through substrate selection and proofreading (editing). During substrate selection, aaRSs discriminate between different tRNAs based on various structural determinants. The major discriminatory determinants are located in the acceptor step of tRNA and the anticodon loop [62]. Because different amino acids can have similar shapes and sizes, discrimination between amino acids is more complex than discrimination between different tRNAs, and requires editing. A ”double sieve” model has been used to explain the high fidelity of aminoacylation [63]. According to this model, the active site of aaRSs serves as a first sieve. This site binds cognate amino-acids, as well as smaller ones, while rejecting amino acids larger than the cognate one. Bound amino acids are then covalently linked with tRNAs. Some aaRSs have domains for editing aminoacyl-tRNAs, which acts as the second sieve by selectively hydrolyzing non-cognate amino-acids from tRNA based on the size and chemical properties of their side chains [63].

Once tRNAs have been charged with amino-acids, they can form ternary complexes with EF-TU and GTP to be delivered to a ribosome. tRNAs can also dissociate from this ternary complex and rebind to aaRSs. For mischarged aa-tRNA, this provides another opportunity for editing. Such *post-transfer editing* in *trans* has been observed for aminoacyl-tRNA synthetases cognate to phenylalanine (PheRS) and proline (ProRS) [64]. Furthermore, different aaRS-like proteins can act as autonomous factors in post-transfer editing. These proteins are similar to editing domains of aaRSs and protect cells against tRNAs charged with non-cognate or even D-amino acids [65, 66]

1.3.4 Phenotypic mutation rates

1.3.4.1 Mistranscription rates

All stages of transcription are subject to noise. The mechanisms and consequences of transcriptional noise have received a lot of attention [6, 20] in studies that focus on stochastic bursts in expression [6, 67, 68]. Transcriptional bursts create heterogeneity in mRNA and protein concentrations in a population, without directly affecting the sequence of protein products.

Some errors of transcription alter the sequence of the mRNA transcript. The first estimates of such mistranscription rates were performed by measuring misreading of nonsense alleles of *lacZ* in *E. coli*. From these measurements, the average mistranscription rate was estimated to be around 10^{-5} per nucleotide [69–71]. With the use of next generation sequencing, it became possible to directly detect mistranscription events occuring between $10^{-4} - 10^{-5}$ per nucleotide in *E. coli* [72]. In *Caenorhabditis elegans*, the average mistranscription rate was estimated to be $\approx 5 \times 10^{-6}$ per nucleotide [73].

1.3.4.2 Mistranslation rates

Ribosomal decoding of mRNA is thought to be the most error-prone step in protein synthesis [22]. The estimated average rate of amino-acid misincorporations is 10^{-3}-10^{-5} per codon in *E. coli* [16, 28] and *S. cerevisiae* [74]. Even higher rates, 10^{-2} per codon, have been observed in some species, such as *Bacillus subtilis* [75]. It is important to note that estimated mistranslation rates vary by more than an order of magnitude across different codons [16, 28]. In other words, average mistranslation rates do not accurately reflect per-codon mistranslations rates. However, specific per-codon mistranslation rates are largely unknown.

Many factors can modulate mistranslation rates. First and foremost, mutations in ribosomal proteins and rRNA can increase or decrease mistranslation rates [27, 76–79]. For example, mutagenesis of ribosomal proteins S4, S5, and S12 has revealed dozens of changes in these proteins that affect mistranslation rates [80–84].

Second, the presence of some antibiotics decreases translational fidelity. The effects of streptomycin, kanamycin, and chloramphenicol on ribosomal fidelity have been well studied [85–88]. In fact, the first mutants affecting translational fidelity were found among *E. coli* colonies resistant to streptomycin [89–91]. Other antibiotics, like kasugamycin, may increase translational fidelity [92].

Third, the quality of the environment is crucial for the accuracy of translation. In particular, it was found that starvation increases mistranslation [93–96]. The temperature and the composition of the growth medium can also affect mistranslation [75, 97].

Fourth, the genetic context (e.g. neighboring codons) can affect mistranslation rates. For example, mistranslation rates might be greatly increased if rare codons occur in tandem [98]. In addition, the same codon can be mistranslated at different rates when found in different locations of the same gene [99, 100]. However, it is not known if this is caused by neighboring codons or other effects, such as mRNA secondary structure.

The final factor in regulating mistranslation is tRNAs. The accuracy of mRNA decoding is in part determined by the competition between cognate and non-cognate tRNAs on the ribosome, and mistranslation rates can be drastically changed by changing tRNA abundance [16]. Even when tRNA concentrations are held constant, chemical modification of anticodon nucleotides can change the specificity and accuracy of decoding [28, 101].

1.3.4.3 tRNA mischarging rates

During translation, the anticodons of aa-tRNAs pair with codons in mRNA. If there is a perfect match between a codon and an anticodon, but if the tRNA was previously charged with an incorrect amino acid, a missense phenotypic mutation will result. In addition to mistranslation,

tRNA mischarging is the most important source of phenotypic mutations, and has been estimated to $\approx 10^{-4}$ per codon [102, 103].

1.3.5 The evolution of phenotypic mutation rates

Phenotypic mutations are at $10^{-3} - 10^{-4}$ per codon orders of magnitude more frequent than genotypic mutations ($10^{-7} - 10^{-11}$ per base pair per generation) in laboratory conditions [16]. Moreover, bacteria from natural isolates tend to have even higher phenotypic mutation rates than laboratory strains [104]. These observations raise the question why phenotypic mutation rates do not evolve to lower values? This is especially puzzling since kinetic proofreading mechanisms that control the accuracy of protein synthesis [56, 57] can in theory support arbitrarily low error rates. Furthermore, it is relatively easy to select for hyperaccurate ribosomes with antibiotics such as streptomycin [91, 105]. The answer probably lies in the cost of increased accuracy [106]. Increased ribosomal accuracy can reduce the rate of translation and energy efficiency [107, 108]. Therefore, translation probably evolved to optimize and balance ribosome speed, efficiency, accuracy, and energetic costs, especially in natural populations [104, 108].

1.4 Evolutionary consequences of phenotypic mutations

1.4.1 Phenotypic mutations and protein evolution

Rates of protein evolution have been studied for a long time. While the same protein can have very similar evolutionary rates in different lineages, different proteins of the same species can evolve at very different rates [109]. Since this variation is often larger than the variation in synonymous substitution rates across the genome, variation in mutation rates among proteins is an unlikely explanation for differences in evolutionary rates. Instead, it is likely that variation in evolutionary rates reflects evolutionary constraints on proteins [110]. Such constraint was thought to be closely related to protein function [111, 112]. However, the quantites that show the strongest influence on the rate of protein evolution are the expression level or the frequency of translation events [109, 113, 114]. Since mistranslation is frequent (see 1.3.4.2) and costly (see 1.3.1), mistranslation-induced costs will increase with translation rates. This led to the hypothesis that mistranslation is the main factor affecting variability in the rate of protein evolution [115]. Theory and experiments demonstrate that mistranslation affects protein evolution in the following ways:

1. *Mistranslation slows the rate of evolution and drives the evolution of translational robustness.* Covariation between protein expression level, codon choice, and evolutionary rates suggest that there is an underlying cost of translation. The so-called mistranslation-induced misfolding hypothesis posits that these costs come from the cytotoxicity of misfolded proteins [115]. In consequence, proteins under a high burden of mistranslation will be under strong selection that favors translationally robust sequences. Translational robustness can be directly affected by stability, which can be increased or decreased by mutations [116, 117]. Indeed, theory and experiments show that protein stability can increase in adaptation to mistranslation (see chapter 2 and [114]). However, because translationally robust sequences are rare in sequence space, mistranslation-induced cost will also constrain evolution [109].

2. *Mistranslation drives the evolution of translational accuracy.* Synonymous codons have different propensity to be mistranslated [16]. Findings that mistranslation is costly suggest that selection should act on the set of optimal codons, and on positions of these codons within genes to increase translational accuracy. Indeed, optimal codons are associated with conserved sites in proteins across different domains of life [115, 118–120]. More specifically, optimal codons are preferentially found at sites where substitutions are most likely to disrupt protein structure [121]. In *E. coli* and *S. dysenteriae* it was found that codon adaptation is even higher in proteins that are not routine clients of chaperones [122]. In other words, selection for translational accuracy is stronger in proteins, where chaperones can not buffer destabilizing effects of mistranslation.

1.4.2 Cellular adaptations to phenotypic mutations

Because phenotypic mutations can be very costly to cells, it is not surprising that cells have evolved a myriad of different mechanisms that increase the accuracy of tRNA charging, transcription, and translation (see 1.3.3). However, these mechanisms do not completely prevent phenotypic mutations, so a set of other mechanisms exists to reduce the severity of mistranslation effects.

The first of these regards the genetic code, which has error-mitigating properties [23]. Specifically, the arrangement of codon-to-amino-acid mappings is efficient at mitigating deleterious effects of random point mutations and phenotypic mutations. That is, the code is organized in such a way that codons that differ by a single nucleotide specify amino acids that are similar in their physicochemical properties. If a codon is misread by the ribosome, such misreading commonly happens through a single mismatch in the codon-anticodon helix [23], and the substituted amino acid will be similar to the original one. Whether or not the structure of the genetic code evolved under selection to minimize the deleterious effects of (phenotypic) mutations [23, 123] is an unresolved question [124].

The second regards nucleotide sequence-level adaptations other than the evolution of high-fidelity synonymous codons at structurally sensitive sites. One of these is an abundance of out-of-frame stop codons. These stop codons are common in prokaryotic genomes, ensuring that nascent erroneous polypeptides are terminated prematurely if ribosomal frameshifts occur during translation [125].

Finally, post-translational processes can increase robustness to phenotypic mutations. Chaperones increase the probability that mutated proteins fold properly [126, 127]. In other words, chaperones can mitigate misfolding as the most deleterious consequence of mistranslation. Also, if other mechanisms have failed, the protein degradation machinery can remove damaged and misfolded proteins [128]. Remarkably, up to 30% of newly synthesized proteins are rapidly degraded, likely because they are erroneously synthesized [129]. As a result of these cellular adaptations, cells can tolerate surprisingly high rates of phenotypic mutations [130–132].

1.4.3 "The look-ahead effect" of phenotypic mutations

Variation is a source of physiological and evolutionary innovations. In populations that are isogenic, phenotypic variation can result from plastic responses to environmental changes or from noise [133]. The idea that such phenotypic plasticity can enhance survival and be selected for dates back to Baldwin[134]. Noise, and phenotypic mutations in particular, could act in the same way. On average, phenotypic mutations will cause a fitness decrease in constant environments. However, phenotypic heterogeneity resulting from statistical protein pools could in principle ensure that a fraction of populations survive an environmental change that would otherwise destroy it. How phenotypic mutations could facilitate adaptation is explained by a hypothesis called "the look-ahead effect" [135].

"The look-ahead effect" is based on a biophysical view of protein evolution. In this view, the evolution of novel enzymatic activities is constrained. Specifically, many mutational paths to novel functions go through intermediates with reduced fitness, and this imposes constraints on the parts of the sequence space that can be explored in a search for functional innovations [136]. Mistranslation could allow selection of low-fitness intermediates of a complex trait that requires multiple mutations [135]. Some of the intermediate mutants will, via errors in translation, be able to synthesize protein pools containing a fraction of the high-fitness protein product, even though the exact DNA sequence for that product is not encoded in the genome. Rare gene sequences that are pre-adapted to accommodate such phenotypic mutations might arise in the process. In other words, the probability of producing high-fitness protein through errors in translation can be selected for. Because mistranslation is generally cytotoxic, the look-ahead effect can influence evolution only under certain environmental conditions. Given strong selection for the optimized

function and large population sizes, even a small fraction of high-fitness proteins will ensure survival, and give the population enough time to genetically assimilate a high-fitness protein through random mutations and fixation [135]. In fact, phenotypic mutations can increase the rate of genomic mutations through a mechanism called translational stress-induced mutagenesis [137]. In other words, phenotypic mutations can not only cause immediate fitness benefits, but they can also increase the adaptive potential of a population by transiently increasing mutation rates [138].

1.5 Methods for measuring mistranslation

Comprehensively measuring mistranslation poses formidable technical problems due to the size of the space of all possible mistranslation events. If we neglect nonsense, frameshifts, and readthrough errors, each sense codon from the genetic code table can in principle be mistranslated into 19 different amino-acids. That means there are 1159 (61x19) possible missense substitutions. Our current estimates are based on only 5% of all possible amino acid misincorporations, and only on few species [16, 22, 28, 74]. Attempts to measure mistranslation rates have been based on many different approaches [22, 132]. Most of the estimates are derived from experiments based on indirect biochemical methods [16, 28]. New experimental methods are sensitive enough to enable detection of mistranslation rates as low as 10^{-6} per codon [28], but comprehensive measurements of mistranslation rates are still laborious and technically challenging. Recently, mass spectrometry (MS) has been used to directly detect mistranslation during the expression of heterologous proteins [97, 98, 139]. I will briefly describe indirect and MS methods.

1.5.1 Detection of an amino acid not usually present in a protein

This strategy depends on detecting the erroneous incorporation of an amino acid that is not found in the error-free variant of the protein. It was used in the first study to measure mistranslation rates, which relied on measuring the misincorporation of radioactively labeled valine into rabbit ovalbumine [25], and estimated the mistranslation rate as $2 - 6 \times 10^{-4}$ per codon. Related approaches rely on detecting changes in protein properties that are caused by amino-acid misincorporation, such as changes in protein mass and charge, using two-dimensional electrophoresis and isoelectric focusing [94, 140].

1.5.2 Mistranslation-induced gain of protein activity

Different gain-of-function reporter genes can be used to measure mistranslation rates indirectly. In this approach, an essential residue of a reporter gene is changed, such that the encoded enzyme is non-functional. Cells transformed with the gene for this enzyme can restore its function through phenotypic mutations. Thus, the activity of the enzyme can be used as a proxy for mistranslation rates. In one study, where a catalytic serine residue in β-lactamase was replaced with a glycine [141], mistranslation was estimated at 10^{-3} per codon. This study also showed that the mistranslation rate depends on which synonymous glycine codon is used. Similarly, restoration of luciferase activity was used to estimate mistranslation rates in *E. coli* [16, 28], *Saccharomyces cerevisiae* [74], and *Mycobacterium smegmatis* [37].

1.5.3 Direct detection of misincorporated amino acids with mass spectrometry

In contrast to indirect biochemical methods, mass spectrometry can be used to directly detect mistranslated proteins. In this approach, large numbers of tandem mass spectra are collected and matched to their corresponding peptides, which are mapped to their source proteins [142]. Peptide variants carrying phenotypic mutations can be detected through shifts in their mass

spectra, relative to their respective wild-type variants. For example, in an important study, tandem mass spectrometry has been used to detect Asn to Asp misincorporation in *E. coli*, which demonstrated that bacterial cells can survive mistranslation rates of up to 10% [131]. In a similar experiment, mistranslation of serine to asparagine was measured in Chinese Hamster Ovary (CHO) and *E. coli* cells during production of monoclonal antibodies [143]. Mistranslation rates were estimated to be on the order of 10^{-3} per codon. The most accurate method based on mass-spectrometry has been used to measure mistranslation during synthesis of recombinant proteins in *E. coli*, monoclonal antibodies in mammalian cells, and human serum albumin [144]. This method allowed rates as low as 10^{-5} per codon to be detected, and revealed that G/U mismatches in codon-anticodon pairing probably contribute the most to mistranslation [144].

1.6 Thesis outline

In my thesis, I studied different aspects of mistranslation, which I now briefly summarize.

In chapter 2 [145], I investigated how proteins adapt to cope with increased rates of mistranslation. Specifically, I evolved an antibiotic resistance protein in *E. coli* hosts with normal and elevated mistranslation rates. In particular, I investigated the interplay between mistranslation and the strength of selection. I found that populations adapt to mistranslation in two ways, depending on the strength of selection. Under stringent selection for protein activity, populations accumulated nonsynonymous changes that increased the stability of proteins, and purged changes that are destabilizing. Under relaxed selection, populations evolved by adopting inefficient initiation codons. This reduced mistranslation-induced costs by lowering the rate of translation.

In chapter 3, I studied how mistranslation affects populations of proteins adapting to a new biochemical activity. To this end, I evolved an antibiotic resistance protein with selection for activity against a new antibiotic. I found that mistranslation does not influence the phenotypic evolution of resistance. While I found no evidence that mistranslation facilitates evolution under directional selection, I found that mistranslating populations accumulate cryptic genetic diversity. I show that this diversity contains protein variants that can facilitate adaptation to other antibiotics.

In chapter 4, I used preexisting mass-spectrometry proteomics data to quantify and characterize mistranslation in two pathogenic bacteria. I found that mistranslation might be more common than previously thought, and that phenotypic mutations can cause radical amino acid changes in proteins. Furthermore, I found that bacteria frequently mistranslate essential proteins. Some of the mistranslated proteins are identically mistranslated in different conditions, and in both bacterial species. This suggests an important role of mistranslation in pathogenic bacteria.

Chapter 2

Mistranslation drives the evolution of robustness in TEM-1 β-lactamase

Published as:
Bratulic S, Gerber F, and Wagner A (2015) *PNAS*, **112**(41):12758-12763

Abstract

How biological systems such as proteins achieve robustness to ubiquitous perturbations is a fundamental biological question. Such perturbations include errors that introduce phenotypic mutations into nascent proteins during the translation of mRNA. These errors are remarkably frequent. They are also costly, because they reduce protein stability and help create toxic misfolded proteins. Adaptive evolution might reduce these costs of protein mistranslation by two principal mechanisms. The first increases the accuracy of translation via synonymous 'high fidelity' codons at especially sensitive sites. The second increases the robustness of proteins to phenotypic errors via amino acids that increase protein stability. To study how these mechanisms are exploited by populations evolving in the laboratory, I evolved the antibiotic resistance gene TEM-1 in *Escherichia coli* hosts with either normal or high rates of mistranslation. I analyzed TEM-1 populations that evolved under relaxed and stringent selection for antibiotic resistance by single molecule real-time sequencing. Under relaxed selection, mistranslating populations reduce mistranslation costs by reducing TEM-1 expression. Under stringent selection, they efficiently purge destabilizing amino acid changes. More importantly, they accumulate stabilizing amino acid changes rather than synonymous changes that increase translational accuracy. In the large populations I study, and on short evolutionary timescales, the path of least resistance in TEM-1 evolution consists of reducing the consequences of translation errors rather than the errors themselves.

2.1 Introduction

Protein synthesis or translation is a key step in genetic information processing. Despite being fundamental to all cellular life, translation is remarkably error-prone. Mistranslation events are estimated to occur once per 10^2-10^4 codons [16, 22, 75]. When mRNA is mistranslated, synthesized proteins carry phenotypic mutations in positions where ribosomes incorrectly decoded the mRNA. A pool of proteins with such phenotypic mutations, also called statistical proteins [23], can differ from error-free proteins in sequence, structure, and function.

Comparative genomics studies show that mistranslation can help explain why highly expressed proteins evolve slowly [109, 113–115]. All else being equal, highly expressed genes experience more translation events, and thus give rise to a higher number of mistranslated

proteins [109, 115]. Mistranslation is costly because it can destabilize proteins, and increases proteotoxic stress by promoting protein misfolding and aggregation [22, 146, 147].

Natural selection can reduce mistranslation costs via two non-mutually exclusive groups of mechanisms. The first increases *translational accuracy*, i.e., it reduces the rate at which translational errors occur. A global increase in accuracy, for example through hyper-accurate ribosomes, affects all proteins but comes with high energetic and kinetic costs [107]. Alternatively, translational accuracy can increase locally at amino acid sites where (phenotypic) mutations would cause the largest fitness defects [121]. This increase is possible through synonymous mutations towards codons with a low propensity for mistranslation [16, 28, 148].

A second group of mechanisms does not reduce mistranslation itself, but mitigates its deleterious consequences, and thus increases the *translational robustness* of proteins. In other words, translational robustness mechanisms mask cytotoxic effects of mistranslation. Some error-mitigation mechanisms are global and affect many proteins. They include chaperones that can help proteins fold even if they harbor destabilizing mutations [149]. In contrast to such global mechanisms, which can become overwhelmed when mistranslation rates are high, local mechanisms are less costly. They rely on (genetic) mutations called suppressors, which increase the stability of a single protein, and thus buffer destabilizing effects of other (phenotypic) mutations [150–152].

The only experimental study of protein evolution under phenotypic mutations focused on errors in transcription [153]. It could not answer how proteins evolve to reduce the mutational load of mistranslation, by increasing their translational accuracy, or by mitigating the effect of errors resulting from such accuracy? Here, I address this question by evolving genes encoding the antibiotic resistance protein TEM-1 β-lactamase in strains of *E. coli* with different rates of mistranslation. I also ask whether relaxed and stringent selection for antibiotic resistance affect the adaptation to elevated mistranslation in different ways. Under relaxed selection, mistranslating populations adapt by reducing TEM-1 expression through inefficient initiation codons, which lowers the cost of mistranslation. Under stringent selection, where reducing gene expression would be detrimental, populations increase translational robustness by accumulating stabilizing and purging destabilizing SNPs.

2.2 Results

To study how proteins adapt to mistranslation in evolving laboratory populations, I experimentally evolved TEM-1 β-lactamase independently in two *E. coli* strains, the wild-type, and the mistranslating, or error-prone *rpsD12* strain (figure 2.1). The error-prone strain carries a mutation in the ribosomal protein S4, which results in increased missense, readthrough, and frameshift (phenotypic) mutations during protein synthesis [16]. Specifically, I evolved TEM-1 in populations of 10^8-10^9 individuals for eight cycles of PCR-based mutagenesis and selection (figure 2.1), with four replicate populations for each of the two host strains and each of the two selection conditions (relaxed: $25\,\mu g\,ml^{-1}$ ampicillin; stringent: $250\,\mu g\,ml^{-1}$ ampicillin). I also evolved two replicate control populations per strain. These populations were mutagenized in the same way as evolved populations, but experienced no selection for β-lactamase activity. I sequenced more than 500 evolved molecules per population using single molecule real-time (SMRT) sequencing [154] (Supplementary table 2.1).

From the sequenced library of ancestral TEM-1 (395 sequences), I estimated the compound sequencing and variant calling error-rate as 4.4×10^{-5} per nucleotide. In control libraries (4567 variants), I observed an average of 0.73 mutations per variant, implying a mutation rate of $\approx 8 \times 10^{-4}$ per nucleotide. Consistent with reports from previous studies [117, 155], my mutagenesis protocol is A→G and T→C biased (Supplementary table 2.2).

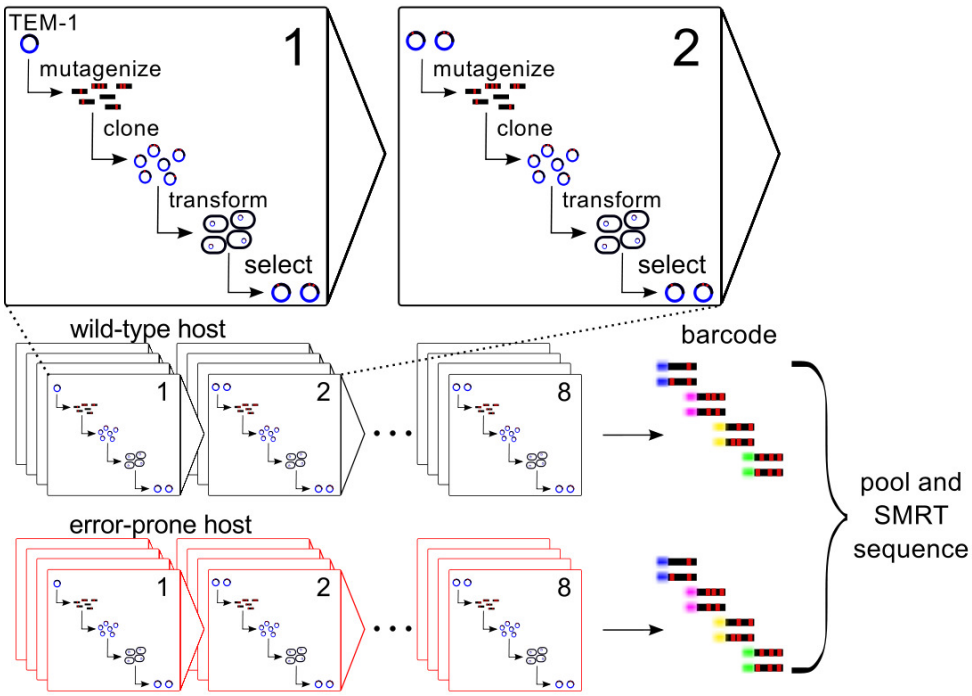

Figure 2.1: Experimental evolution of TEM-1 under mistranslation. In each round of evolution, I exposed TEM-1 to mutagenic PCR and recloned the resulting mutant alleles into a fresh plasmids backbone, thus ensuring that only the TEM-1 evolves in my experiments. I transformed plasmids with mutagenized TEM-1 into host cells (wild-type or error-prone), and exerted relaxed and stringent selection for antibiotic resistance by growing 10^8-10^9 transformed hosts in liquid LB media with ampicillin (25 or 250 $\mu g\,ml^{-1}$, respectively). Subsequently, I isolated plasmids and used them as templates for the next round of evolution. I evolved 4 replicate populations per host and per selection regime, for a total of 16 populations. After 8 cycles of evolution, I subjected evolved TEM-1 populations to single-molecule real time (SMRT) sequencing.

2.2.1 Mistranslation slows down TEM-1 evolution

The TEM-1 protein has two parts: the N-terminal signal peptide (the first 25 residues in Ambler numbering [156]), and the mature enzyme. The signal peptide guides the translocation of TEM-1 to the periplasmic space. Once translocation is complete, the signal sequence is cleaved and the mature TEM-1 folds into its active conformation. The signal peptide controls the expression, and the localization of TEM-1. I observe that SNPs in the signal peptide have frequencies up to $\approx 26\,\%$ (Supplementary figure 2.6 and supplementary table 2.7). In contrast, SNPs in the mature part of TEM-1 all have frequencies below $5\,\%$ (Supplementary figure 2.6, supplementary tables 2.5 and 2.6).

Next, I compared the average number of SNPs per TEM-1 variant among the two strains. Wild-type populations have a higher relative frequency of variants with more SNPs than error-prone populations (figure 2.2A), and selection makes this difference more pronounced. Specifically, under relaxed selection, wild-type and error-prone populations have different estimated means of 5.41 and 5.08 SNPs per variant, respectively (estimates and comparisons are based on general linear models (GLMs) with a Wald test of the corresponding estimate, $z = 5.90, P < 0.001$). Under stringent selection the estimated mean number of SNPs per variant was also different between wild-type (5.28) and error-prone (4.59) populations (Wald test, GLM, $z = 11.98, P < 0.001$).

The accumulation of nonsynonymous SNPs in mature TEM-1 proteins (figure 2.2B in black) shows that error-prone populations are under stronger purifying selection. Specifically, under relaxed selection, error-prone populations accumulated significantly fewer nonsynonymous SNPs per variant than wild-type populations (1.76 vs. 2.01 nonsynonymous

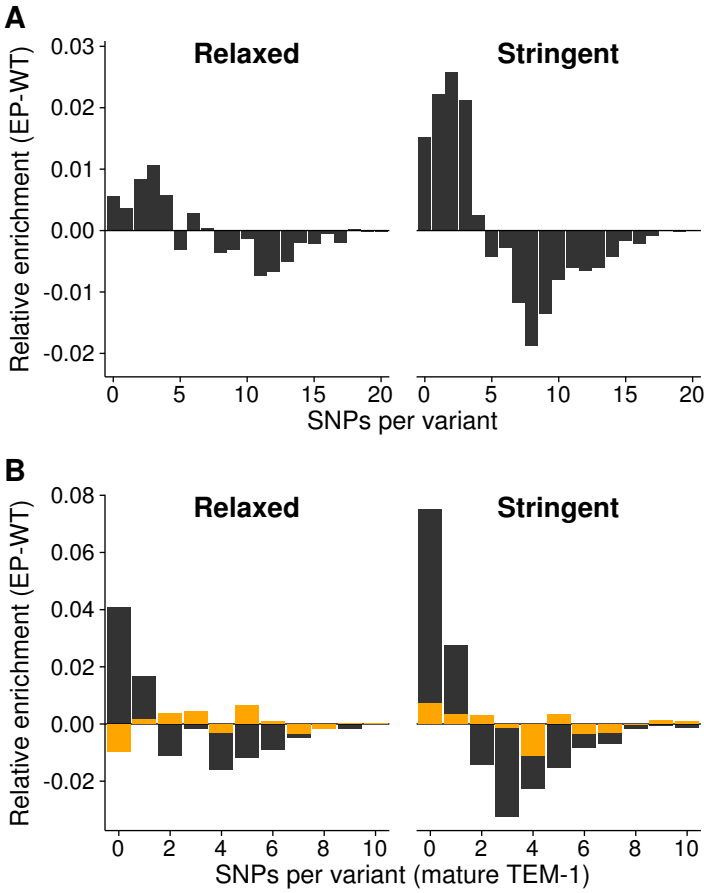

Figure 2.2: SNP enrichment in evolved populations for relaxed and stringent selection. (A) The relative enrichment of variants with a given number of SNPs (horizontal axis) for the whole protein, calculated by subtracting the frequency (Supplementary figure 2.7) of variants with this number of SNPs in the wild-type host from that in the error-prone host. The data show that error prone strains have more variants with fewer SNPs, and fewer variants with more SNPs, relative to the wild-type. For this analysis, I binned variants according to the number of observed SNPs, and calculated the frequency of alleles in a bin relative to the pooled data from all replicates for a given host. (B) as in (A), but only for synonymous (orange) and nonsynonymous (black) SNPs in the mature part of TEM-1.

SNPs per variant; Wald test, GLM, $z = 8.81, P < 0.001$). Under stringent selection, this difference became even more pronounced (1.51 vs. 1.90 nonsynonymous SNPs per variant; Wald test, GLM, $z = 13.81, P < 0.001$). Additionally, the ratio of nonsynonymous to synonymous SNPs was significantly lower in error-prone lines, indicating stronger purifying selection under both selection regimes (relaxed selection: 0.83 vs. 0.95 for mistranslating and for wild-type populations, respectively; Wald test, GLM, $z = 8.20, P < 0.001$; stringent selection: 0.71 vs. 0.87; Wald test, GLM, $z = 11.4, P < 0.001$). Further evidence for stronger selection under mistranslation comes from the distribution of fitness effects of nonsynonymous SNPs, where error-prone populations are depleted of deleterious, and enriched in neutral and beneficial nonsynonymous SNPs (Supplementary figure 2.9).

As opposed to these indicators of purifying selection, the mean number of synonymous SNPs per variant did not differ significantly between error-prone and wild-type populations (figure 2.2B in orange, relaxed selection: 2.12 vs. 2.14 syn. SNPs, Wald test, GLM, $z = -0.51, P \approx 1$; stringent selection: 2.18 vs. 2.13, Wald test, GLM, $z = 1.39, P = 0.32$). Taken together, my observations show that the nonsynonynmous SNPs I observe are subject to purifying selection associated with error-prone translation. In contrast, synonymous SNPs are accumulating neutrally with respect to mistranslation, contrary to what one would expect if there was strong selection for increased translational accuracy through high-fidelity

synonymous SNPs in my experimental system.

2.2.2 TEM-1 adapts to mistranslation through increased stability and changes in expression

To see how mistranslation affects the robustness of evolved proteins, I first predicted the stability effects of observed SNPs using FoldX [157]. FoldX can compute the thermodynamic impact of a mutation, expressed as $\Delta\Delta G$, where a mutation with $\Delta\Delta G<0$ is stabilizing. Error-prone populations accumulated more stabilizing SNPs and fewer destabilizing SNPs, a difference that is once again more pronounced under stringent selection (figure 2.3A; Wilcoxon rank-sum test, two-sided $P < 0.001$ for both relaxed and stringent selection).

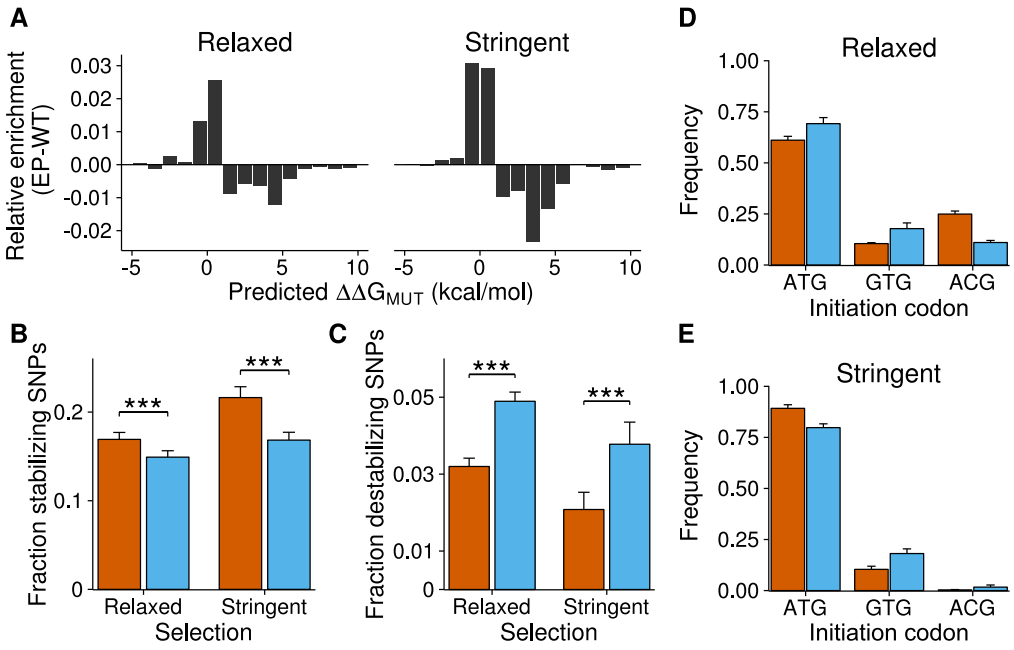

Figure 2.3: Stability and expression changes in evolved populations. (A) Thermodynamic impact of enriched SNPs as predicted by FoldX [157]. SNPs are binned according to $\Delta\Delta G$. Positive values of enrichment (vertical axis) correspond to a larger number of mutations with a given $\Delta\Delta G$ in error-prone populations. (B) Fraction of experimentally validated stabilizing SNPs among all nonsynonymous SNPs. (C) Fraction of validated destabilizing SNPs among all nonsynonymous SNPs. (D) Frequencies of sequences with each of three initiation codons (horizontal axis) under relaxed selection. (E) Like (D), but for stringent selection. Only codons that appear with a frequency greater than 5 % in at least one of the populations are included in (D) and (E). Error-bars represent standard deviations across four replicate populations.

To validate computational predictions of FoldX, I compiled a list of mutations known from experiment to either increase or decrease the stability of TEM-1 [117, 150–152, 158–165]. TEM-1 in error-prone populations accumulated significantly more stabilizing SNPs than in wild-type populations, and it did so for both selection regimes (figure 2.3B; relaxed selection: 17.1% vs 14.8%, Wald test, GLM, $z = -5.02, P < 0.001$; stringent selection: 21.6% vs. 16.7%, Wald test, GLM, $z = -9.17, P < 0.001$). At the same time, error-prone populations accumulated fewer destabilizing SNPs compared to wild-type populations (figure 2.3C; relaxed selection: 3.2% for error-prone, 4.8% for wild-type, Wald test, GLM, $z = 6.67, P < 0.001$; stringent selection: 2.1% for error-prone populations, 3.9 % for wild-type populations, Wald test, GLM, $z = 7.61, P < 0.001$). Furthermore, the well-known stabilizing mutation M182T is the nonsynonymous SNP with the highest frequency in two populations under mistranslation and stringent selection (Supplementary table 2.5). No significant differences exist in the accumulation of synonymous SNPs at sites where mutations are known to affect stability (Supplementary figure 2.10).

SNPs found at the initiation codon (within the signaling peptide), have the highest frequencies in my dataset (figure 2.3D and E, Supplementary table 2.7). Under relaxed selection, sequences evolved in error-prone hosts are more likely to have non-ATG initiation codons (figure 2.3D, 38.6% vs 31.1%, Wald test, GLM, $z = -9.27, P < 0.001$), which reduce efficiency of translation initiation [166, 167]. In contrast, under stringent selection, sequences evolved in wild-type hosts are more likely to have non-ATG initiation codons (figure 2.3E, 20.0% vs 10.7%, Wald test, GLM, $z = 14.21, P < 0.001$).

2.2.3 Mistranslating populations accumulate nonsynonymous SNPs in surface residues

I next examined where in the TEM-1 tertiary structure SNPs accumulated during evolution (figure 2.4). In general, mutations that affect a protein's core tend to be more destabilizing than mutations of surface residues [168]. Core residues are buried and have a low solvent accessible surface area (SASA), while surface residues have high SASA. I computed SASA for each of the residues affected by SNPs, and found that residues with higher SASA tend to be enriched in nonsynonymous SNPs in error-prone populations, relative to wild-type populations (relaxed selection: Wilcoxon rank-sum test, two-sided $P < 0.001$; stringent selection: Wilcoxon rank-sum test, two-sided $P < 0.001$, figure 2.4B and figure 2.4C).

In contrast, distributions of SASA for residues harboring synonymous SNPs did not differ between error-prone and wild-type strains (relaxed selection: Wilcoxon rank-sum test, two-sided $P \approx 1$; stringent selection: Wilcoxon rank-sum test, two-sided $P = 0.56$, figure 2.4D).

2.3 Discussion

I find first that mistranslation can indeed affect the rate of protein evolution. Specifically, error-prone proteins accumulate fewer non-synonymous changes, and show a lower ratio of non-synonymous to synonymous changes. This pattern of evolution is readily explained through the observation that most nonsynonymous mutations destabilize proteins [116, 168]. In populations subject to high rates of mistranslation, proteins harboring such destabilizing mutations suffer additional destabilizing effects from phenotypic mutations, and thus become even more destabilized. In such populations, a greater fraction of nonsynonymous mutations should thus get eliminated by natural selection, just as I observed. Also consistent with my observation is that those nonsynonymous changes that survive high rates of mistranslation occur preferentially on the TEM-1 surface (figure 2.4B and figure 2.4C) where they are less likely to be destabilizing [168]. Second, several lines of evidence show that cells adapt to mistranslation by reducing TEM-1 mistranslation costs (figure 2.5). They do so with two different strategies, depending on whether selection for antibiotic resistance is relaxed or stringent.

Under relaxed selection (low ampicillin concentrations), error-prone populations reduce TEM-1 expression by adopting inefficient non-ATG initiation codons, which reduce the cost of misfolding by reducing TEM-1 expression (figure 2.5). Changes in initiation codons have the highest frequency among all SNPs I observed (figure 2.3D and E, Supplementary table 2.7). Such changes are known to reduce TEM-1 expression and with it the cost of mistranslation. Specifically, the GTG initiation codon, which is used in about 14% of *E. coli* genes [169] has a 1.5-3 times lower initiation efficiency than ATG [166]. Similarly, ACG can serve as an initiation codon [169], but its initiation efficiency is only 1-3% of that of ATG [167]. In addition, both GTG and ACG initiation codons are frequently observed in comprehensive TEM-1 mutagenesis libraries selected at low levels of ampicillin [170]. Reducing the concentration of TEM-1 is a simple strategy to mitigate mistranslation costs, but it is only viable where amounts of ampicillin are so low that TEM-1 expression can be reduced without adverse effects.

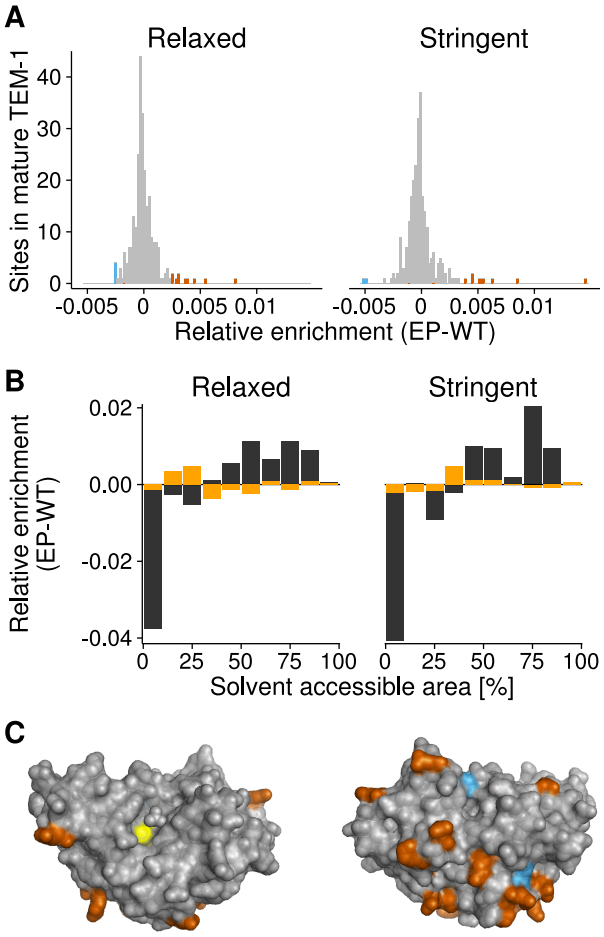

Figure 2.4: Accumulating SNPs and where they occur on the TEM-1 structure. (A) Histogram of sites that harbor host-specific enriched nonsynonymous SNPs. Strongly enriched bins (more than two SD) are shown in red (error-prone) and blue (wild-type). (B) Relative enrichment of synonymous (orange) and nonsynonymous (black) SNPs in areas with a given solvent accessibility (horizontal axis). (C) TEM-1 structure with the sites of polymorphic nonsynonymous SNPs colored as in (A) according to their relative enrichment. The active site is shown in yellow. Left and right structures are rotated horizontally by 180 degrees relative to each other.

In contrast, under stringent selection (high antibiotic concentration), a high active concentration of TEM-1 is needed to degrade ampicillin. In this condition, mistranslating lines, where a fraction of functional TEM-1 is already lost to mistranslation, should not be able to reduce TEM-1 expression much further. This prediction is borne out by my observation that error-prone lines are more likely to use the standard ATG initiation codon under stringent selection. A related observation was made in experiments with elevated mistranscription rates [153], where transcription with error-prone RNA polymerase reduces the effective expression of TEM-1, and populations with increased mistranscription adapted to higher concentrations of ampicillin by increasing TEM-1 expression.

Error-prone populations subject to stringent selection (high ampicillin concentration) mitigate the effects of mistranslation not by reducing TEM-1 expression, but by accumulating stabilizing and depleting destabilizing mutations in TEM-1 (figures 2.3A and B). Remarkably, the change with the highest frequency in my mistranslating populations is the well-known M182T substitution (Supplementary table 2.5). Frequently observed in natural TEM-1 isolates and in laboratory evolution experiments [163], M182T increases the stability of TEM-1, making it more robust to genetic mutation and denaturation [116]. In addition, error-prone populations are impoverished relative to wild-type in predicted and known destabilizing SNPs in TEM-1 (figure 2.3A and C). Furthermore, they have especially few nonsynonymous SNPs in

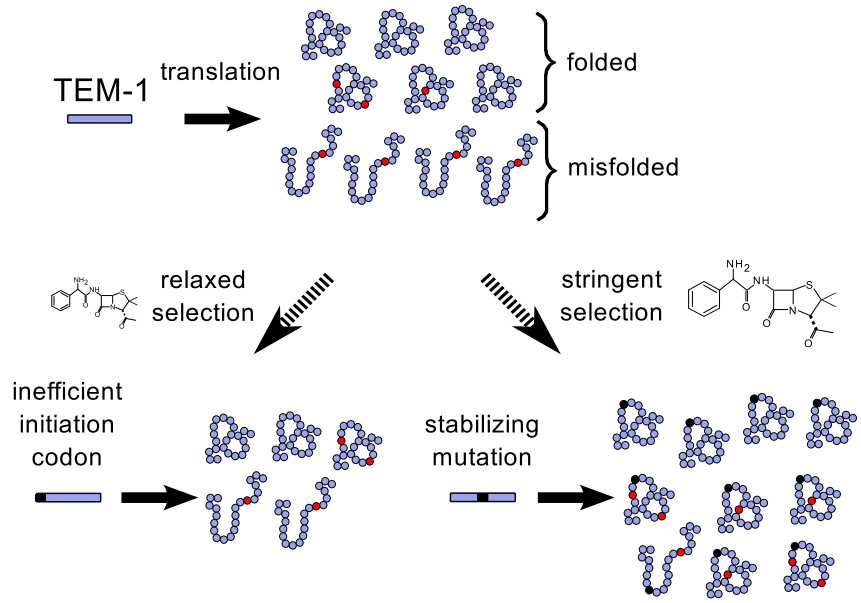

Figure 2.5: A model for the evolution of translational robustness under error-prone translation. Phenotypic mutations are shown in red, genotypic mutations in black. The strength of selection affects adaptation. Under relaxed selection (low ampicillin concentration), only small amounts of β-lactamase are needed, and populations reduce the cost of mistranslation by reducing expression. Under stringent selection (high ampicillin concentration), larger amounts are needed, and populations reduce the cost of misfolding by accumulating stabilizing changes and purging destabilizing changes.

the TEM-1 core, where amino acid changes would be strongly destabilizing (figure 2.4). In other words, mistranslation causes efficient purging of destabilizing mutations.

Taken together, I find that under under laboratory conditions evolution adapts to mistranslation by mitigating the damage it causes. My observations are consistent with previous experiments showing that increased mistranscription rates [153] can lead to increases in protein stability [165].

My third observation pertains to whether evolution alters the robustness or the accuracy of translation. Specifically, does TEM-1 evolve increased translational accuracy, which can occur by synonymous changes towards high-fidelity codons? My observations suggest that, at least in my experimental system, the answer is no. First, synonymous SNPs do not generally accumulate at a higher rate in error-prone populations. Second, the incidence of synonymous SNPs at sites where mutations are known to have stabilizing effects is not greater in error prone populations. Third, the incidence of synonymous SNPs at sites where mutations have destabilizing effects is not lower in error-prone populations. Finally, the incidence of synonymous SNPs in codons adjacent to those with known stability effects does not differ between error-prone and wild-type populations (Supplementary figure 2.10). These findings are consistent with theoretical predictions that adaptation to mistranslation may predominantly occur through increased translational robustness because robustness provides bigger benefits and is thus easier to evolve [114, 115]. They are also rendered plausible by two further observations. First, the number of known stabilizing and destabilizing amino acid changes is large (Supplementary Methods 2.7.13), which implies that evolutionary modulation of protein stability is easily achieved. Second, once a stabilizing SNP reduces the destabilizing effects of mistranslation, further selection for 'high fidelity' synonymous SNPs will be less effective [171].

The conditions of my experimental evolution differ from those experienced by many natural populations. For example, my experimental design imposed strong selection (high antibiotic concentrations), a high mutation rate, and large populations, as well as few ($\approx$ 50) cell generations. The last condition is especially important, because the evolution of synonymous changes may require many more generations [172]. These differences may help explain why

selection for translational accuracy can be effective in some natural populations [115, 118, 120, 173], but was not effective in my experiments.

Error-prone protein translation has occurred since life's earliest days [23], and it has contributed to the evolution of a robust genetic code [123]. My observations demonstrate that it can still influence the structure of modern proteins. The path of least evolutionary resistance, in laboratory evolved TEM-1, reduces the consequences of errors rather than the errors themselves.

2.4 Materials and methods

For detailed description of experimental procedures, see section 2.7.

2.4.1 Strains and plasmids

The wild-type and the error-prone hosts were derived from *E. coli* strain MG1655, and were isogenic except for the rpsD allele. That is, the error-prone host carried an *rpsD12* allele [30], while the wild-type had a normal rpsD allele. I used the high-copy number plasmid pHS13T (Supplementary figure 2.11B), derived from pHSG396 [174], which carries a chloramphenicol resistance marker, as the vector for TEM-1 evolution.

2.4.2 Directed evolution

To mutagenize the TEM-1 population, I used error-prone PCR with nucleoside analogues [155]. After PCR, I digested, purified, and ligated the mutagenized TEM-1 sequences into fresh plasmid backbones. Subsequently, I transformed the ligation product into electrocompetent DH5α cells to ensure plasmid methylation, which resulted in library sizes between 10^5-10^6 sequences. After recovering the transformed cells, I grew them overnight in LB media supplemented with 34 μg/mL of chloramphenicol, purified plasmids (preselection libraries) from these overnight cultures, and transformed these libraries into electrocompetent *rpsD12* or wt cells (library sizes: 10^8-10^9 sequences). I allowed recovered transformants to grow for approximately six generations in LB media supplemented with 34 μg/mL chloramphenicol (for plasmid maintenance), as well as either 25 μg/mL or 250 μg/mL of ampicillin, for relaxed and stringent selection regime, respectively. After purifying plasmids from the resulting postselection libraries, I used them as templates in the next round of evolution, as well as for single molecule real-time sequencing.

2.4.3 Primary data analysis

I assembled consensus reads (referred to as variants in the manuscript) of TEM-1 sequences from subreads (≈13.5 passes per consensus TEM-1 variant) with the SMRTAnalysis v2.3 package. I mapped reads to the reference (ancestral) TEM-1 sequence using BLASR [175], and filtered mapped reads that spanned the entire TEM-1 coding region to an average Phred quality above 20. I considered a mismatch of a TEM-1 sequence read to the reference TEM-1 sequence a true SNPs only if its Phred quality score was above 20.

2.4.4 Statistical methods

Unless specified otherwise, I used generalized linear models (GLMs) [176] to compare the four groups given by the selection regimes (relaxed, stringent) and the host strains (wild-type, error-prone). I report the estimated means of Quasi-Poisson models for count data (number of

SNPs) and estimated proportions (such as dN/dS ratios) of Quasi-Binomial models. For comparisons involving GLMs, I indicate the z-value of the corresponding Wald test statistic and the corresponding P-value, which I adjusted for multiple testing with the Holm-Bonferroni procedure. I took the grouping of the data in four replicate populations into account via an extension to generalized linear mixed models [177]. However, based on model diagnostics, I decided to report the estimates of the GLMs. Further details on the statistical methods are given in Supplementary Methods 2.7.12.

2.5 Supplementary figures

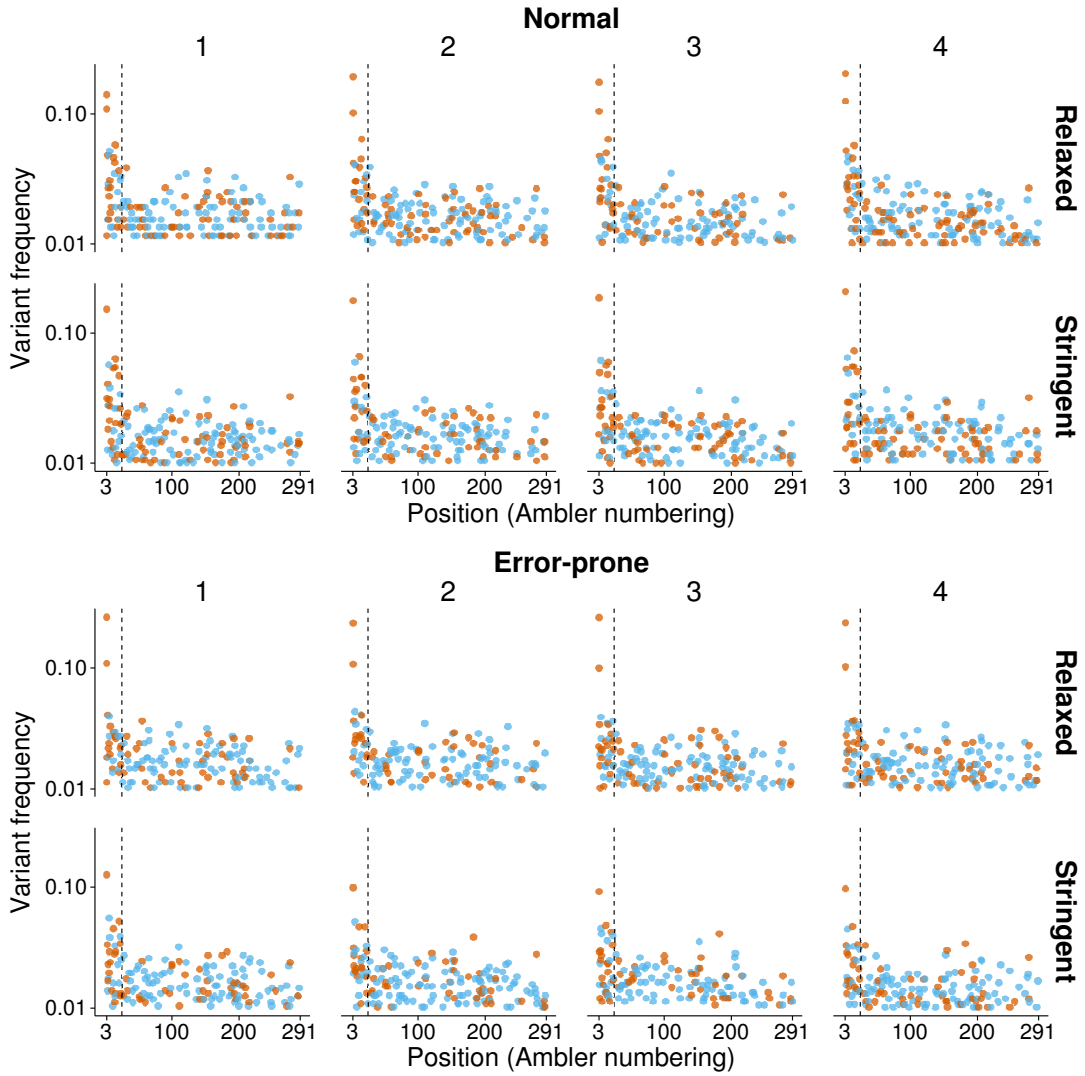

Figure 2.6: The distribution of SNP frequencies along the TEM-1 coding gene for all 16 evolved populations (4 populations per strain and per treatment). Synonymous SNPs are shown in red, nonsynonymous SNPs in blue. The dashed line indicates the boundary between the signal and the mature part of the TEM-1 enzyme. For clarity, only variants with frequencies exceeding one percent are shown.

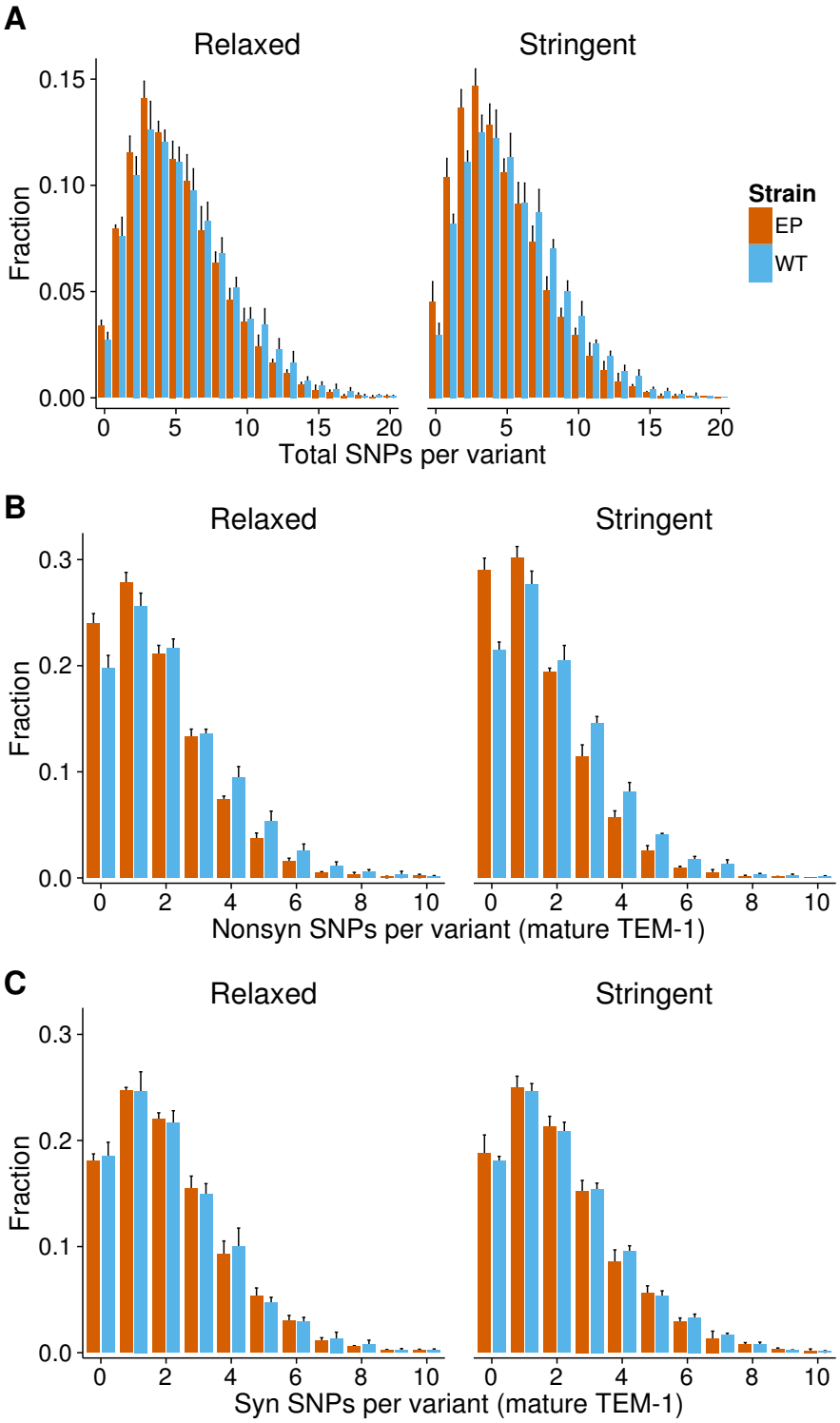

Figure 2.7: The enrichment of variants with a given number of SNPs (horizontal axis). (A) Total number of SNPs in the TEM-1 cds. (B) Total number of nonsynonymous SNPs per gene (mature TEM-1). (C) Total number of synonymous SNPs per gene (mature TEM-1). Error-bars show standard deviations for the 4 replicate populations.

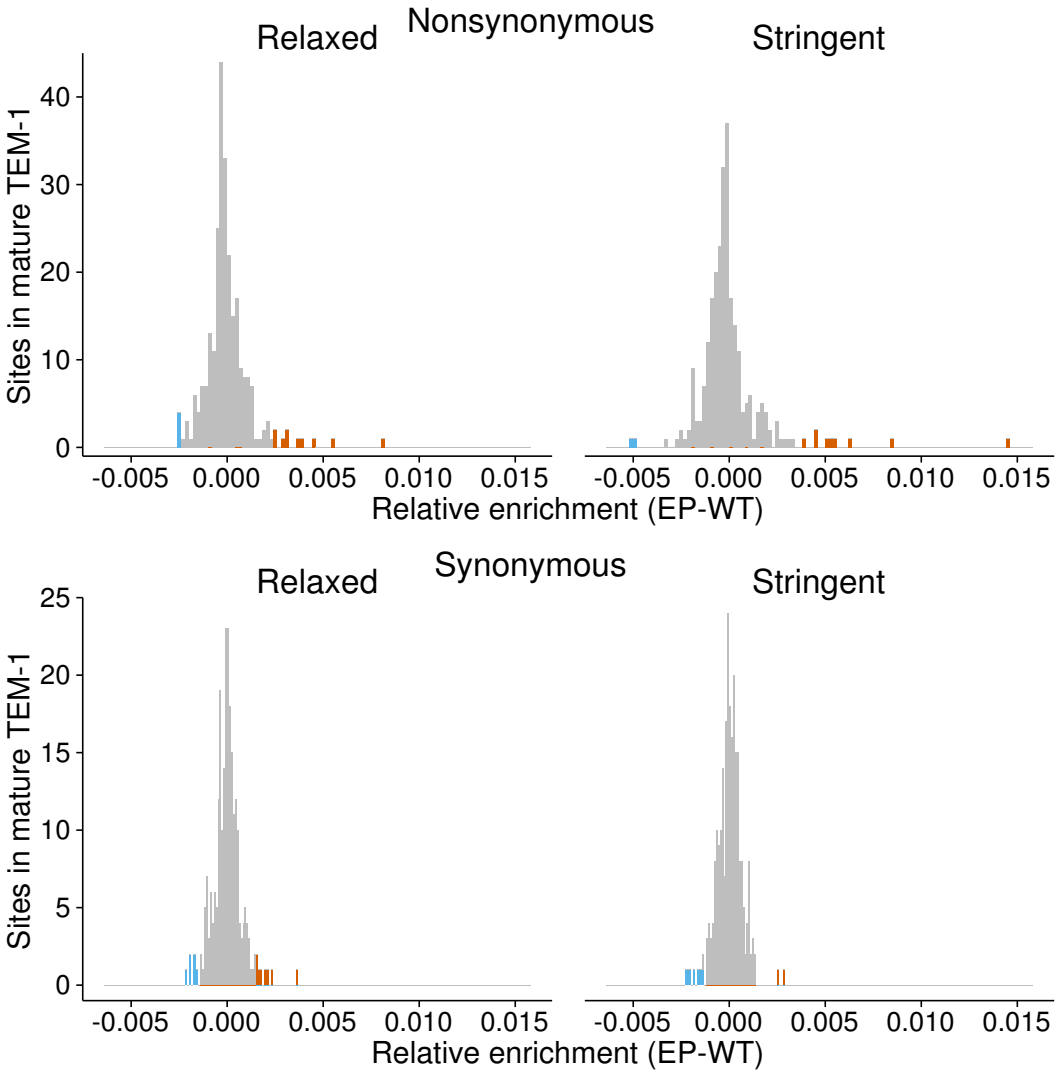

Figure 2.8: The number of sites that harbor host-specific enriched SNPs. Strongly enriched bins (more than two SD) are shown in red (error-prone) and blue (wild-type). (A) Sites enriched with nonsynonymous SNPs. (B) Sites enriched with synonymous SNPs. The distribution of enrichment values for nonsynonymous SNPs is much broader, indicating stronger selection acting on nonsynonymous than synonymous changes.

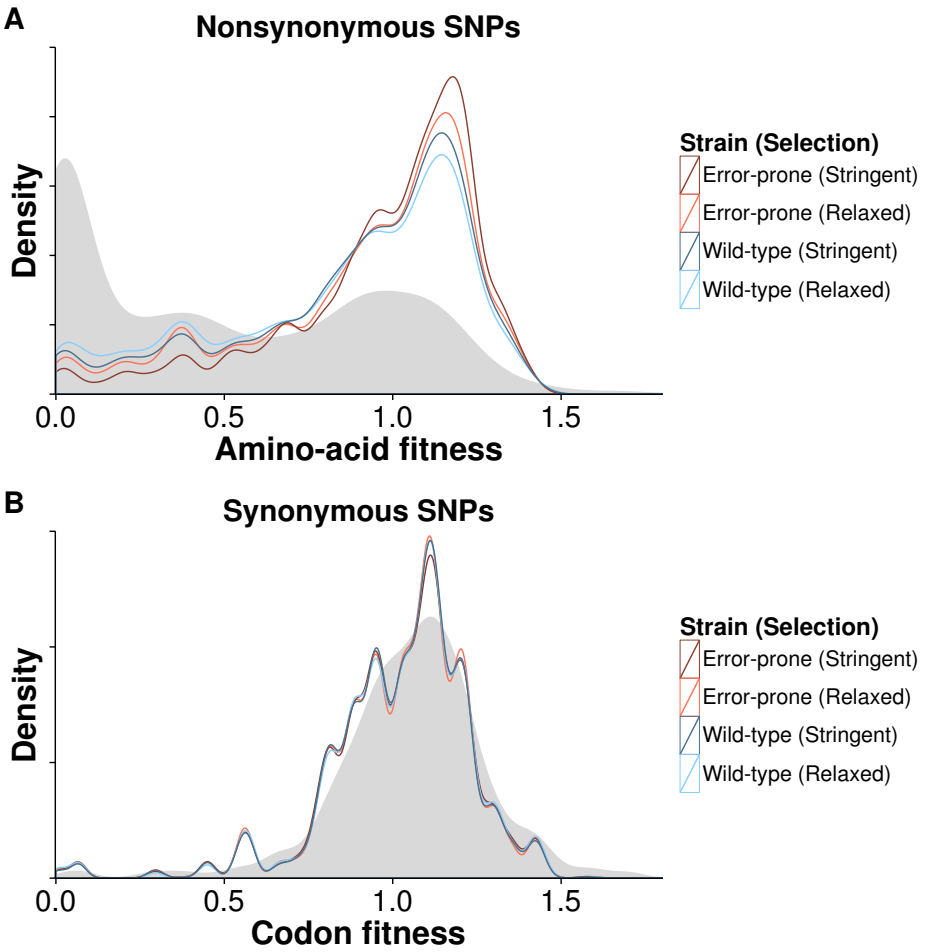

Figure 2.9: Distribution of fitness effects (DFE) for SNPs in evolved populations. Each SNP was assigned a fitness value taken from experimental data in [170]. (A) DFE for nonsynonymous SNPs. Grey density curve shows the DFE of all nonsynonymous mutations from the published dataset [170]. Each of the colored density curves (see the legend) is based on pooled data from four replicate evolved populations. These curves show that error-prone populations experience more efficient purging of deleterious nonsynonymous SNPs, and enrichment in neutral and beneficial SNPs. (B) DFE for synonymous SNPs. Grey density curve shows the DFE of all synonymous mutations from the published dataset [170].

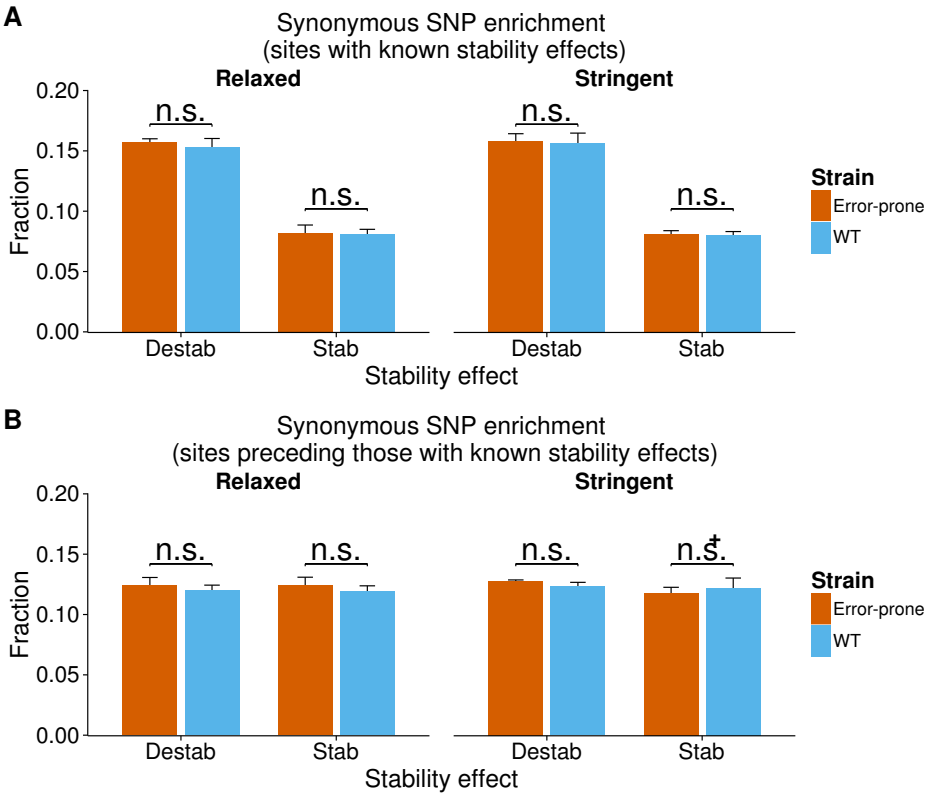

Figure 2.10: Enrichment of synonymous SNPs at sites with experimentally validated (de)stabilizing effects (see 2.7.13). Sites were classified as stabilizing, destabilizing, or unknown. (A) Fraction of synonymous SNPs at sites with (de)stabiling effects. (B) Fraction of synonymous SNPs at sites preceding sites with (de)stabiling effects (-1 codon relative to sites shown in (A)). Differences in enrichment between error-prone and wild-type populations are nonsignificant in all of the eight pairwise comparisons ($P = 0.498$ for the comparison marked with $+$, $P \approx 1$ for all other comparisons).

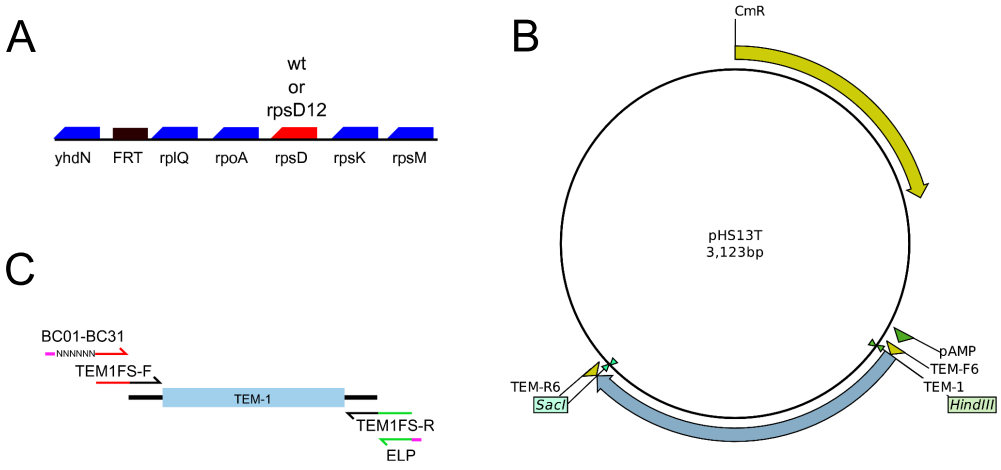

Figure 2.11: (A) The genomic region where the *rpsD* gene is located. FRT scar is located downstream of the *rplQ* gene(B) The pHS13T plasmid used for in the evolution experiment. Chloramphenicol resistance gene is shown in yellow, while the TEM-1 is show in blue. TEM-1 is under the control of a constitutive promoter (green) from pBR322. HindIII and SacI restriction sites flank the TEM-1 coding region. TEM-F6 and TEM-R6 are binding sites for primers used in mutagenic PCR. (C) Primers used in a 2-step PCR for uniquely barcoding populations.

2.6 Supplementary tables

Table 2.1: Sequencing and SNP statistics. Population names are given as Host_Selection_Replicate. The number of SNPs refers to all observed SNPs before (raw), and the number of SNPs after quality filtering (HQ).

Population	Sequences	Mean Quality	SNPs (raw)	SNPs (HQ)	SNPs per variant (raw)	SNPs per variant (HQ)
WT_Relaxed_1	521	40.1	4598	2931	8.83	5.63
WT_Relaxed_2	1956	39.9	17413	10936	8.90	5.59
WT_Relaxed_3	2434	39.8	20913	12588	8.59	5.17
WT_Relaxed_4	1381	39.9	12289	7819	8.90	5.66
WT_Stringent_1	1980	39.9	16721	10487	8.44	5.30
WT_Stringent_2	1439	39.9	11632	7537	8.08	5.24
WT_Stringent_3	2288	40.0	19042	12168	8.32	5.32
WT_Stringent_4	852	39.9	7186	4644	8.43	5.45
Error-prone_Relaxed_1	879	39.9	7210	4491	8.20	5.11
Error-prone_Relaxed_2	1921	40.0	15724	9783	8.19	5.09
Error-prone_Relaxed_3	2950	39.9	23889	15119	8.10	5.13
Error-prone_Relaxed_4	2129	40.0	17422	10866	8.18	5.10
Error-prone_Stringent_1	1430	40.0	10902	6845	7.62	4.79
Error-prone_Stringent_2	1650	40.0	12214	7977	7.40	4.83
Error-prone_Stringent_3	1407	40.1	10133	6403	7.20	4.55
Error-prone_Stringent_4	1665	40.1	11798	7149	7.09	4.29
WT_Control_1	1395	40.5	1719	949	1.23	0.68
WT_Control_2	2161	40.6	2686	1529	1.24	0.71
Error-prone_Control_1	564	40.7	748	420	1.33	0.74
Error-prone_Control_2	412	40.7	581	336	1.41	0.82
TEM-1(Ancestor)	395	40.8	38	15	0.10	0.04

Table 2.2: Mutation rate matrix, calculated from pooled sequence data of non-selected control libraries. Rates are expressed per site and per cycle of mutagenesis

→	A	C	G	T
A	9.9858×10^{-01}	5.5411×10^{-05}	1.3506×10^{-03}	1.0884×10^{-05}
C	1.9565×10^{-05}	9.9978×10^{-01}	5.4348×10^{-06}	1.9239×10^{-04}
G	2.0376×10^{-04}	3.9757×10^{-06}	9.9978×10^{-01}	1.4909×10^{-05}
T	9.3234×10^{-06}	1.3498×10^{-03}	6.8371×10^{-05}	9.9857×10^{-01}

Table 2.3: GLM estimates with 95% confidence intervals - Relaxed selection

Estimate	WT	WT(CI95%)	EP	EP(CI95%)	z	P	$P_{adjusted}$
Total SNPs	5.41	5.32-5.49	5.08	5.00-5.15	5.90	3.731×10^{-09}	1.492×10^{-08}
Total nonsyn SNPs (Mature TEM-1)	2.01	1.97-2.05	1.76	1.73-1.80	8.81	1.457×10^{-18}	1.312×10^{-17}
Total syn SNPs (Mature TEM-1)	2.12	2.08-2.17	2.14	2.10-2.18	-0.51	6.075×10^{-01}	9.999×10^{-01}
dN/dS	0.95	0.93-0.97	0.83	0.81-0.84	8.20	2.309×10^{-16}	1.616×10^{-15}
Frac. stabilizing SNPs [%]	14.85	14.23-15.47	17.12	16.49-17.74	-5.02	5.152×10^{-07}	1.545×10^{-06}
Frac. destabilizing SNPs [%]	4.82	4.44-5.19	3.21	2.91-3.50	6.67	2.589×10^{-11}	1.295×10^{-10}
Frac. non-ATG init. codons [%]	31.13	29.99-32.28	38.63	37.56-39.71	-9.27	1.908×10^{-20}	1.908×10^{-19}

Table 2.4: GLM estimates with 95% confidence intervals - Stringent selection

Estimate	WT	WT(CI95%)	EP	EP(CI95%)	z	P	$P_{adjusted}$
Total SNPs	5.28	5.19-5.36	4.59	4.51-4.67	11.98	6.946×10^{-33}	5.557×10^{-32}
Total nonsyn SNPs (Mature TEM-1)	1.90	1.86-1.94	1.51	1.47-1.54	13.81	4.801×10^{-43}	4.320×10^{-42}
Total syn SNPs (Mature TEM-1)	2.18	2.13-2.22	2.13	2.09-2.18	1.39	1.643×10^{-01}	3.286×10^{-01}
dN/dS	0.87	0.85-0.89	0.71	0.69-0.73	11.45	2.242×10^{-30}	1.570×10^{-29}
Frac. stabilizing SNPs [%]	16.72	16.06-17.37	21.64	20.81-22.48	-9.17	4.953×10^{-20}	2.476×10^{-19}
Frac. destabilizing SNPs [%]	3.91	3.57-4.25	2.06	1.77-2.35	7.61	2.748×10^{-14}	8.243×10^{-14}
Frac. non-ATG init. codons [%]	19.99	19.02-20.96	10.73	9.96-11.5	14.21	7.541×10^{-46}	7.541×10^{-45}

Table 2.5: The ten nonsynonymous SNPs with the highest frequency in the dataset that affect the mature TEM-1. The position is given in Ambler numbering [156]. Reference and SNP refer to the amino-acid found in the ancestral TEM-1 and the evolved population, respectively. The frequency is the ratio between the number of times a SNP is observed and the total number of sequences from the population

Position	Reference	SNP	Times observed	Total sequences	Frequency	Population
182	M	T	58	1407	4.12	EP_Stringent_3
182	M	T	65	1650	3.94	EP_Stringent_2
32	K	R	20	521	3.84	WT_Relaxed_1
154	N	S	19	521	3.65	WT_Relaxed_1
56	I	V	32	879	3.64	EP_Relaxed_1
182	M	T	57	1665	3.42	EP_Stringent_4
32	K	R	55	1665	3.30	EP_Stringent_4
276	N	D	65	1980	3.28	WT_Stringent_1
276	N	D	17	521	3.26	WT_Relaxed_1
276	N	D	27	852	3.17	WT_Stringent_4

Table 2.6: The ten synonymous SNPs with the highest frequency in the dataset that affect the mature TEM-1. The position is given in Ambler numbering [156]. Reference and SNP refer to the codon found in the ancestral TEM-1 and the evolved population, respectively. The frequency is the ratio between the number of times a SNP is observed and the total number of sequences from the population

Position	Reference	SNP	Times observed	Total sequences	Frequency	Population
28	GAA	GAG	76	1956	3.89	WT_Relaxed_2
64	GAA	GAG	31	852	3.64	WT_Stringent_4
152	TTG	CTG	82	2288	3.58	WT_Stringent_3
152	TTG	CTG	50	1407	3.55	EP_Stringent_3
110	GAA	GAG	70	1980	3.54	WT_Stringent_1
28	GAA	GAG	30	852	3.52	WT_Stringent_4
110	GAA	GAG	85	2434	3.49	WT_Relaxed_3
110	GAA	GAG	67	1921	3.49	EP_Relaxed_2
121	GAA	GAG	18	521	3.45	WT_Relaxed_1
110	GAA	GAG	30	879	3.41	EP_Relaxed_1

Table 2.7: Variants from the signal peptide with the highest frequency in evolved populations. The position is given in Ambler numbering [156]. Reference and SNP refer to the codon found in the ancestral TEM-1 and the evolved population, respectively. The frequency is the ratio between the number of times a SNP is observed and the total number of sequences in the population. Position 3 in Ambler numbering corresponds to the initiation codon of TEM-1.

Position	Reference	SNP	Times observed	Total sequences	Frequency	Population
3	ATG	ACG	231	879	26.28	EP_Relaxed_1
3	ATG	ACG	772	2950	26.17	EP_Relaxed_3
3	ATG	ACG	505	2129	23.72	EP_Relaxed_4
3	ATG	ACG	453	1921	23.58	EP_Relaxed_2
3	ATG	GTG	177	852	20.77	WT_Stringent_4
3	ATG	GTG	282	1381	20.42	WT_Relaxed_4
3	ATG	GTG	377	1956	19.27	WT_Relaxed_2
3	ATG	GTG	428	2288	18.71	WT_Stringent_3
3	ATG	GTG	256	1439	17.79	WT_Stringent_2
3	ATG	GTG	427	2434	17.54	WT_Relaxed_3
3	ATG	GTG	302	1980	15.25	WT_Stringent_1
3	ATG	GTG	73	521	14.01	WT_Relaxed_1
3	ATG	GTG	181	1430	12.66	EP_Stringent_1
3	ATG	ACG	172	1381	12.45	WT_Relaxed_4
3	ATG	ACG	57	521	10.94	WT_Relaxed_1
3	ATG	GTG	96	879	10.92	EP_Relaxed_1
3	ATG	GTG	207	1921	10.78	EP_Relaxed_2
3	ATG	ACG	255	2434	10.48	WT_Relaxed_3
3	ATG	GTG	219	2129	10.29	EP_Relaxed_4
3	ATG	ACG	199	1956	10.17	WT_Relaxed_2

Table 2.8: Primers and barcodes used for mutagenesis and sequencing

Primer	Sequence	Barcode
BC01	GGTAGGAGCAATGTAAAACGACGGCCAGT	AGCAAT
BC02	GGTAGGCCTGTTGTAAAACGACGGCCAGT	CCTGTT
BC03	GGTAGGGGGTTTGTAAAACGACGGCCAGT	GGGTTT
BC04	GGTAGGGAAGGCGTAAAACGACGGCCAGT	GAAGGC
BC05	GGTAGGATCTCAGTAAAACGACGGCCAGT	ATCTCA
BC06	GGTAGGATGGATGTAAAACGACGGCCAGT	ATGGAT
BC07	GGTAGGATGTCTGTAAAACGACGGCCAGT	ATGTCT
BC08	GGTAGGCGTGACGTAAAACGACGGCCAGT	CGTGAC
BC17	GGTAGGCGATGCGTAAAACGACGGCCAGT	CGATGC
BC18	GGTAGGGATAGCGTAAAACGACGGCCAGT	GATAGC
BC19	GGTAGGGTCAGAGTAAAACGACGGCCAGT	GTCAGA
BC20	GGTAGGTTAAGCGTAAAACGACGGCCAGT	TTAAGC
BC21	GGTAGGAACCTGGTAAAACGACGGCCAGT	AACCTG
BC22	GGTAGGCTTTGCGTAAAACGACGGCCAGT	CTTTGC
BC23	GGTAGGTGGAGAGTAAAACGACGGCCAGT	TGGAGA
BC24	GGTAGGAATTGTGTAAAACGACGGCCAGT	AATTGT
BC25	GGTAGGTGACGAGTAAAACGACGGCCAGT	TGACGA
BC26	GGTAGGCAAATAGTAAAACGACGGCCAGT	CAAATA
BC29	GGTAGGGTTGGGGTAAAACGACGGCCAGT	GTTGGG
BC30	GGTAGGGCTTAGGTAAAACGACGGCCAGT	GCTTAG
BC31	GGTAGGTAGCCAGTAAAACGACGGCCAGT	TAGCCA
TEM1FS-F	GTAAAACGACGGCCAGTGAATAATATTGAAAAAGGAAGC	-
TEM1FS-R	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGTAAACTTGGTCTGACAGGAGC	-
ELP	GGTAGGCAAGCAGAAGACGGCAT	-
TEM-F6	GCTTAAGAATAATATTGAAAAAGG	-
TEM-R6	GAATTGTAAACTTGGTCTGACA	-

2.7 Supplementary methods

2.7.1 Media and antibiotics

I used Difco LB broth (BD) for growth and selection of all strains. For preparing competent cells, I used SOB media (Sigma). For recovery after electroporation I used SOC media, which I prepared by adding 20 mM glucose to SOB media. For antibiotic selection, I used kanamycin sulfate at 50 μg/L, chloramphenicol at 25 and 34 μg/l, ampicillin sodium salt (Sigma) at 25, 100, and 250 μg/L. I used saline (0.9 μg/L NaCl) to prepare serial dilutions for library size estimations.

2.7.2 Strains

I used the *E. coli* strain DH5α for all cloning steps, including the preparation of TEM-1 libraries prior to selection in each cycle of the evolution experiments. The ribosomal mutant with increased mistranslation rate, *rpsD12*, was kindly provided by T. Nyström [30]. In order to minimize the probability of background mutations that could affect protein evolution, I transfered the rpsD12 allele into a fresh genetic background. To this end, I first used PCR-based recombineering [178] to integrate a kanamycin resistance cassette flanked by FRT sites [178] into the genome of the *rpsD12* strain. The integration site was downstream of the rpoA operon. I isolated the genomic DNA from this construct, and used Phusion (Thermo Scientific) PCR to amplify the region spanning the mutation and the resistance cassette. I then used recombineering to integrate the PCR-amplified fragment into a clean MG1655 background (CGSC#7740). To avoid nonspecific mutations that might result from recombineering, I used P1 transduction to transfer the mutation linked to the kanamycin resistance cassettes into the MG1655 background. Finally, I removed the KanR cassette by transforming transductants with a flipase plasmid pCP20 [179]. I induced the flipase by growing transformed cells at 37°C, and plating them on nonselective LB agar plates. I picked clones that were sensitive to kanamycin (resistance cassette was excised, leaving an FRT scar behind [178]) and ampicillin (temperature sensitive pCP20 plasmid was lost). I confirmed the final construct (figure 2.11A) by colony PCR and Sanger sequencing.

To control for the presence of the FRT scar in the rpsD12, I repeated the same cloning procedure with wild-type MG1655 strain. The final construct (referred to as wild-type or normal strain in the rest of the manuscript) had a wild-type rpsD allele, and an FRT site at the same location as the error-prone rpsD12 strain (figure 2.11A).

2.7.3 Plasmids

I used a high copy number plasmid with a chloramphenicol resistance marker, based on pHSG396 [174], as a backbone for cloning and evolving TEM-1. First, I removed the lac promoter from the plasmid using Phusion PCR. Next, I introduced the coding region of TEM-1 with its constitutive promoter (pAMP) from pBR322 into the modified pHSG396 plasmid. The coding region of TEM-1 was flanked by SacI and HindIII restriction sites. I called this plasmid pHS13T (figure 2.11B). To facilitate gel extraction of vector backbones for recloning, I constructed a second vector, pHS13K. This vector differed from pHS13T by having a KanR cassette from pKD4 [178] as a filler sequence between SacI and HindIII sites. A majority of positions (176 out of 272) in the ancestral TEM-1 sequence have non-optimal codons, based on codon optimality indices in [121].

2.7.4 Electrocompetent cells

To ensure high and reproducible transformation efficiency across all strains, I used electroporation in all our transformations. I prepared electrocompetent cells by glycerol/mannitol density step centrifugation [180]. In short, I diluted 2 mL of an overnight culture in 200 mL SOB media. I incubated this culture with shaking (250 rpm) in a 2 L shake

flask at 37°C, until the OD_{600} reached the values of 0.4-0.6. I chilled the culture in iced water for 15 min, and centrifuged at 1500 g and 4°C, for 15 min in a 5810-R Eppendorf centrifuge with the F-34-6-38 rotor. I resuspended the pellet in 40 mL of ice-cold ddH_2O, and split it into two 50 mL tubes. I slowly added 10 mL of ice-cold 20% (w/v) glycerol + 1.5 % (w/v) mannitol to the bottom of each tube with a 12 mL pipette. I centrifuged the suspension again at 1500 g at 4°C for 15 min, with acceleration/deceleration set to zero. I removed the supernatant by aspiration, and resuspended the pellet in 1 mL of ice-cold 20% (w/v) glycerol + 1.5 % (w/v) mannitol. I aliquoted the cell suspension (80 μL for DH5α, and 50 μL for *rpsD12* and wt strains) in 1.5 mL tubes, froze them in a dry ice-ethanol bath, and stored in -80°C.

2.7.5 Mutagenesis

To introduce genetic diversity into TEM-1 populations, I used a 25 cycle error-prone PCR with nucleoside analogues [155]. A 100 μL PCR reaction contained 10 ng of the template plasmid (pHS13T in the first round, and selected plasmid population in subsequent rounds), with 400 μM dNTPs (Thermo Scientific), 2.5 U Taq polymerase (NEB), Thermopol buffer (NEB), 3 μM 8-oxo-GTP and 3 μM dPTP (Trilink Biotechnologies), 400 nM of primers TEM1-F6 and TEM-R6 (table 2.8). To remove the template plasmid, I treated the PCR product with the restriction enzyme DpnI for 2 h at 37°C. Subsequently, I inactivated all enzymes by adding 0.6 U of proteinase K (Thermo Scientific) and incubated for 1h at 50°C, followed by a 15 min proteinase K inactivation at 80°C.

2.7.6 Library cloning

I carried out the double restriction of the mutagenized TEM-1 pool with 20 U of SacI-HF and HindIII-HF (NEB) for 2h at 37°C, followed by 20 minutes at 80°C. I then purified double digested inserts with the QIAprep PCR purification kit (Qiagen) and eluted them in 2.5 mM Tris-Cl, pH 8.5. In parallel, I double digested the plasmid backbone by incubating pHS13K with 20 U of SacI-HF and HindIII-HF for 16 h. I gel purified the digested vector and dephosphorylated it by incubating with 5 U of Antarctic Phosphatase (NEB) for 1 h, followed by a 20 minute inactivation at 80°C. I then ligated 19 ng of insert (TEM-1 pool) and 50 ng of digested and dephosphorylated vector in 20 μL reactions with 10 U of T4 DNA ligase (NEB) for 16h at 4°C. I inactivated the T4 DNA ligase by incubating for 10 min at 65°C. I precipitated the ligation product by adding 80μL of H_2O, 20 μg of glycogen (Thermo Scientific), 50 μL of 7.5 M ammonium acetate (Sigma), and 2.5 volumes of ice-cold ethanol. I incubated the mixture at -80°C for 20 min, centrifuged for 20 min at 18000 g, washed in 800 uL of 70% cold ethanol, centrifuged and washed again. I dried the pellet under vacuum for 15 min, and then resuspended in 15 uL of 2.5 mM Tris-Cl, pH 8.5.

2.7.7 Preselection libraries

Because I derived the wild-type and the *rpsD12* strain from a restriction-positive MG1655 strain, direct transformation of non-methylated ligation products would result in low transformation efficiency due to restriction. To ensure plasmid methylation before selection in wild-type and *rpsD12* strains, I transformed ligation products into restriction-deficient DH5α cells. To this end, I mixed 80 μL of electrocompetent DH5α cells with 4 μL of the precipitated ligation product, and electroporated using a Micropulser electroporator (Bio-Rad) set on EC3 (15 kV/cm) and 0.2 cm electroporation cuvettes (Cell Projects). Immediately after electroporation, I added 1 mL of pre-warmed SOC media to transformed cells, and transfered the suspension to a 24-well plate. I allowed cells to recover by incubating the plate at 37°C with shaking at 400 rpm for 1.5 h. After the recovery period, I centrifuged the plate and aspirated the supernatant from the plate. I resuspended the cell pellet in 5 mL of LB media supplemented with 34 μg/mL of chloramphenicol. I used a 50 μL cell suspension aliquot to estimate library size by making serial

dilutions in saline and plating on LB agar plates with 20 μg/mL chloramphenicol. Through this procedure, I estimated library sizes to lie between 10^5-10^6 sequences. I incubated transformed cells overnight at 37°C with shaking at 320 rpm. The next morning, I stored 1 ml of the overnight culture as a glycerol stock, and used the rest to purify plasmids with a QIAPrep miniprep kit (Qiagen).

2.7.8 Selection

I transformed 50 μL of electrocompetent *rpsD12* or wt cells with 1 μL of purified preselection libraries. The electroporation and recovery procedures were the same as for preselection libraries. After 1.5 h of recovery in 1 mL SOC media, I prepared serial dilutions from 50 μL of cell suspension, and plated the dilution on LB agar plates with 20 μg/mL chloramphenicol to estimate the library size. Through this procedure, I estimated library sizes to lie between 10^8-10^9 sequences. I centrifuged the remaining recovered cell suspension for 15 min at 2000 g, and resuspended cell pellets in 3 mL LB with 34 μg/mL chloramphenicol, and 25 μg/mL, or 250 μg/mL of ampicillin, for relaxed and stringent selection, respectively. Selection lasted for approximately 6 generations (2:07 h for wild-type, 4:23 h for *rpsD12*), after which I isolated plasmids using the QIAPrep miniprep kit (Qiagen).

2.7.9 Control libraries

To estimate mutation rates in each mutagenesis cycle, I constructed two control libraries per host strain. These libraries were subject to the same procedure as libraries under selection, except that the selection media contained only 34 μg/mL chloramphenicol and no ampicillin. I subjected these control libraries to a single round of evolution.

2.7.10 Sequencing library preparation and SMRT sequencing

I amplified libraries from the last (8th) round of selection in a two step PCR (figure 2.11C). In the first step, I used a 25-cycle PCR with the Phusion polymerase to amplify the coding region of TEM-1 with TEM1FS-F and TEM1FS-R primers (table 2.8). I gel purified PCR products and used them as templates for a second 25-cycle PCR with barcoded primers BCXX and ELP (see table 2.8 for primer sequences). I used 6 bp-long barcodes described in [181]. I purified PCR products from the second PCR using a QIAprep PCR purification kit (Qiagen). To check the quality and concentrations of amplicons in each library, I used the Agilent 2200 TapeStation System (Agilent Technologies). To account for sequencing and library preparation errors, I amplified and barcoded an additional library from an ancestral TEM-1 sequence. Finally, I combined 20 ng of DNA from each library to create a final amplicon pool for sequencing. I produced a SMRTbell library from the amplicon pool with the DNA Template Prep Kit 2.0 (250bp - 3Kb) (Pacific Biosciences p/n 001-540-726). To this end, I inspected amplicon size and integrity on an Agilent Bioanalyzer 2100 1Kb DNA Chip. I then used polishing enzymes to end-repair 500-750 ng of DNA from the amplicon pool.Subsequently, I created the SMRTbell template by blunt end adapter ligation. After that,I used the Agilent Bioanalyzer 12Kb DNA Chip and a Qubit Fluorimeter (Life technologies) to confirm the quality of DNA in the library and estimate its concentration. Finally, I used a DNA/Polymerase P4 binding kit (Pacific Biosciences) according to the manufacturer instructions to create a ready-to-sequence SMRTbell-Polymerase Complex. I programmed the Pacific Biosciences RS2 instrument to sequence the library on 2 SMRT cells v3.0 (Pacific Biosciences), using P4/C2 chemistry, the magnetic bead loading method, and taking 2 movies of 180 minutes. After the run, I generated a sequencing report via the SMRTportal, in order to assess adapter dimer contamination, sample loading efficiency, average read-length, and the number of filtered sub-reads.

2.7.11 Primary data analysis

I assembled consensus reads of TEM-1 variants (reads of insert) from subreads using the SMRTAnalysis v2.3 package. I filtered reads of insert according to a) the minimum number of full pass subreads (4), b) the minimum predicted consensus accuracy (0.9), and c) read of insert length (850-1200 bp). With a mean number of ≈13.5 passes per read of insert, this procedure resulted in 51,555 reads, with a mean read length of 979 bp, and an average read quality of ≈0.98. I mapped reads to the reference (ancestral) TEM-1 sequence using BLASR [175] with a minimum accuracy of 0.9, and a minimum mapped length of 850 bp. The resulting total number of mapped reads was 51,365, with the average mapped read length being 973 bp. The mean mapped subread concordance was 0.97 I further filtered mapped reads to include only those reads with average Phred quality > 20, and spanning the entire coding region of the TEM-1 reference in the alignment. I demultiplexed the filtered set of reads according to their barcodes, using custom Python scripts based on the *pbcore* module (`http://github.com/PacificBiosciences/pbcore`). The final set of reads (32,032 sequences) contained only sequences whose barcodes perfectly matched those I used during library preparation. Because indels are a major source of errors for SMRT sequencing, and because more than 98 % of indels in TEM-1 are loss-of-function mutations [170], I focused our analysis on point mutations. I considered a mismatch of a TEM-1 sequence read to the reference TEM-1 sequence a true SNPs only if its Phred quality score was above 20 (see table 2.1 for summary statistics). I repeated all the analyses with the Phred quality score filter of 0, 10, 30 and 40, but this did not change any of the results. A small fraction (less than 1%) of sequences lacked a stop codon, or had an internal stop codon. I excluded these sequences from the analysis.

2.7.12 Statistical methods

The experiment had a 2×2 design featuring the grouping factors selection (relaxed, stringent) and strain (wild-type, error-prone). I performed four replicate evolution experiments (four independently evolving replicate populations) for each of the four factor combinations. Unless specified otherwise, I used generalized linear models (GLMs) [176] to compare the four groups given by the selection regimes (relaxed, stringent) and the host strains (wild-type, error-prone). I report the estimated means of Quasi-Poisson models for count data (number of SNPs) and estimated proportions (such as dN/dS ratios) of Quasi-Binomial models. For comparisons involving GLMs, I indicate the z-value of the corresponding Wald test statistic and the corresponding P-value, which I adjusted for multiple testing with the Holm-Bonferroni procedure.

I took the grouping of the data in four replicate populations into account via an extension to generalized linear mixed models [177]. In all GLMMs the estimated variance of the random effect was very low and the changes in the fixed effect estimates were negligible compared to the corresponding GLM estimates. In some cases the GLMM fitting procedure did not converge or estimated zero variance of the random effect. For these reasons, and to be consistent throughout the manuscript, I used the estimates of the GLMs (for 95% confidence intervals see tables 2.3 and 2.4). Where explicitly stated, I performed additional comparisons using a two-sided Wilcoxon rank-sum test

Since I performed multiple tests based on the same subset of data, I used the Holm-Bonferroni procedure to keep the family-wise error rate below 5%.

I used the software R (version 3.1.3) for all statistical analyses, and the R-package lme4 to fit GLMM models [182].

2.7.13 Stabilizing and destabilizing mutations

To study the stability effect of mutations in evolved populations, I compiled a list of mutations known to increase the stability of TEM-1. This list included known global suppressors [150–152, 158–164], and mutations enriched in TEM-1 stabilization studies [117, 165]. Specifically, the set of stabilizing mutations included the following changes: V31R, D35Q, I47V, L51F, L51I, L51P, N52S, N52D, N52T, N52H, F60Y, P62S, E63A, E63G, E63K, E63D, G78A, V80I, S82H, Q90H, G92D, N100D, K111R, K111Q, K111T, R120G, R120G, E147G, E147A, E147K, H153R, M182V, M182T, M182R, M182K, M182L, A184V, T188I, L201P, L201Q, L201R, I208L, I208V, I208M, I208T, A224V, A224T, E240K, E240G, E240H, R241H, I247V, T265M, R275Q, R275L, N276D, K288R, K288E, and K288T. Similarly, I compiled a list of destabilizing mutations from the same set of publications. This set included mutations experimentally shown to affect stability, as well as mutations purged from TEM-1 libraries in stabilization studies [117, 165]. The set of destabilizing mutations included: P27S, P27L, Y46H, Y46D, Y46C, D50G, D50E, D50N, L57F, L57A, L57R, L57P, F66L, F66S, F66V, F66C, M68V, M68L, M68T, L76N, L76S, L102T, L102S, L102V, L102W, Y105C, Y105W, Y105H, S106P, S106L, L138V, L138F, L138P, L138R, L152S, L152V, L152T, D163G, D163A, D163E, D163N, R164S, R164C, D179G, D179N, D179V, M186T, M186V, M186L, M211V, M211T, M211R, M211L, S223W, G238S, I263S, I263T, I263L, Y264N, Y264H, Y264C, Y264S, E37G, E48G, V74G, R83C, E89G, S130N, D131A, N136S, D157G, L169P, N170H, T180A, A187T, L199P, Q206P, W210R, M211I, L221R, R222C, K234Q, G238D, and I246N.

2.7.14 Solvent accessible surface area and stability effect calculations

I calculated the solvent accessible surface area (SASA) for all residues in mature TEM-1 using DSSP 2.0.4 [183] based on the PDB structure 1XPB. I normalized computed SASA values for all residues based on the empirical normalization values according to the Table 1 in [184]. I assigned a SASA value to each nonsynonymous mutation in our dataset based on the residue it affected. I used FoldX [157] to calculate stability effects ($\Delta\Delta G_{MUT}$) for SNPs observed in evolved populations. I used the PBD structure 1XPB for these calculations, according to a published procedure [168]. In short, I first used the repair function of FoldX to optimize the PDB file for mutant stability calculation. Next, I built models of single mutants from nonsynonymous mutations observed in our dataset. Finally, I computed the stability effect of a mutation as the difference in energies between the mutant and the wild-type structures. i.e. $\Delta\Delta G_{MUT} = \Delta G_{MUT} - \Delta G_{WT}$. I binned calculated $\Delta\Delta G_{MUT}$ into 1 kcal/mol bins, and assigned all values > 10 kcal/mol to the 9-10 kcal/mol bin.

Chapter 3

Mistranslation increases genetic diversity under directional selection

Abstract

How new phenotypes evolve is one of the central questions of evolutionary biology. Mutational pathways leading to innovations often go through regions of sequence space with low fitness (fitness valleys). Phenotypic mutations have been proposed as one of the mechanisms that might enable populations to cross fitness valleys. Phenotypic mutations appear when ribosomes mistranslate mRNA, and introduce errors into nascent polypeptides. On the one hand, such phenotypic errors reduce fitness because they destabilize proteins and promote misfolding. On the other hand, they also lead to an increase in proteomic diversity. Recent studies hint at a potential of phenotypic mutations to confer new biochemical activities. Even a small fraction of protein variants with new activity synthesized through mistranslation can help a population to survive an environmental challenge. A surviving population will then have an opportunity to genetically assimilate the new protein variant through random mutation and fixation. This theoretical proposition is called "the look-ahead effect" of phenotypic mutations. To validate this hypothesis, I designed an *in vitro* experiment based on the evolution of the antibiotic resistance gene TEM-1. I evolved TEM-1 in *Escherichia coli* hosts with either wild-type or mistranslating ribosomes. Specifically, I subjected four independent populations of TEM-1 per host to four rounds of evolution subject to increasing concentrations of cefotaxime. I analyzed all populations by single molecule real-time sequencing. I found no evidence of "the look-ahead effect" during the evolution of resistance to cefotaxime. First, I found that the mistranslation does not affect the phenotypic evolution, as measured by the increase in resistance against cefotaxime relative to ancestral TEM-1. Second, mistranslating populations fix fewer SNPs and diverge slower. Third, mistranslation causes higher repeatability in genetic composition of populations at the end of the experiment. However, mistranslation leads to an accumulation of cryptic genetic diversity. I show that more diverse populations grow to higher population densities under low and intermediate concentrations of antibiotics that were not previously encountered during evolution. My results demonstrate that even in the absence of a direct "look-ahead effect", phenotypic mutations can increase cryptic genetic variation which can be important in adaptating to unpredictable environments.

3.1 Introduction

3.1.1 Molecular noise and mistranslation

Molecular infidelities govern all biochemical processes, from DNA replication [185, 186] to metabolism [187, 188] and gene regulation [189]. Protein synthesis is no different. Even though accurate decoding of genetic information is critical for protein function, translation can be

surprisingly error-prone [16, 22, 75]. Its accuracy can be compromised during transcription [15, 190], tRNA amino-acylation [61], and translation itself [16]. Of these processes, translation is the most error-prone, because ribosomes incorrectly decode mRNA, thus creating missense, readthrough, or frameshift errors at high rates [22]. These errors, sometimes called phenotypic mutations, can constrain protein evolution and compromise protein stability [22, 115]. However, evidence is mounting that phenotypic mutations can also have benefits across all domains of life [35, 37, 39, 40, 42].

3.1.2 Mistranslation and evolution

Because phenotypic errors can be costly, cells have evolved a wide range of error-reduction and error-mitigation mechanisms [191, 192]. In addition to driving the evolution of cellular responses, mistranslation can affect the evolution of proteins in a direct way. How mistranslation affects proteins evolving under purifying selection has been studied through theory and experiments, leading to four major findings. First, because phenotypic mutations essentially destabilize proteins, mistranslation imposes strong constraints on protein stability, and slows down the rate of evolution (chapter 2, [115]). Second, destabilizing effects of mistranslation can drive the evolution of protein robustness and stability (chapter 2, [114, 153]). Third, proteins can locally reduce mistranslation rates by adopting high-fidelity codons (synonymous changes) at structurally sensitive sites [118, 120, 121]. Fourth, mistranslation can influence the evolution of gene expression. More specifically, when functional proteins are lost due to mistranslation, expression can increase to compensate for that loss [153]. In contrast, if proteins are expressed gratuitously, expression can be lowered to reduce the costs of mistranslation-induced misfolding (chapter 2).

3.1.3 "The look-ahead effect" of phenotypic mutations

Mistranslation is a consequence of biophysical constraints and limits of translational accuracy, hence its origins are clearly non-adaptive [23]. However, traits or processes which are non-adaptive can be coopted or exapted for a novel use [193]. A recent hypothesis proposing the exaptation of mistranslation is "the look-ahead effect" [135]. When a population experiences an environmental change, many individuals will be poorly adapted. In the search for a new fitness peak, many mutational paths go through intermediates with reduced fitness, and this imposes constraints on the parts of the sequence space that can be explored by the population [136]. The structure of the genetic code determines which amino acids substitutions can occur through single nucleotide changes, thus imposing additional constraints on accessible mutational paths [194]. Alternative genetic codes or mistranslation might allow populations to explore different evolutionary paths, or even reach different fitness peaks [195].

Mistranslation could in principle change accessible mutational pathways by allowing low-fitness intermediates to survive. The reason is that phenotypic mutations can lead to the synthesis of a high-fitness protein from a low fitness genotype, and thus facilitate the evolution of a trait that requires more than one mutation [135]. Given strong selection and large population sizes, even a small fraction of high-fitness protein may ensure a cell's survival, and give a population enough time to genetically assimilate the high-fitness protein through random mutations and fixation. A scenario, where phenotypic mutations (stop codon readthrough) serve as stepping stones for evolution was predicted by theory [196], and found in nature for the evolution of intracellular compartmentalization in metabolic enzymes [197]. However, the evolutionary potential of phenotypic missense mutations has not been experimentally studied.

3.1.4 An experimental test of "the look-ahead effect"'

I sought to experimentally test the "look-ahead effect" and to study how mistranslation affects the evolution of protein populations adapting to new biochemical activities. To this end, I evolved an antibiotic resistance enzyme independently under normal and elevated mistranslation rates. I used TEM-1 β-lactamase, an enzyme optimized to hydrolyze ampicillin, and experimentally evolved it towards activity against a new substrate, cefotaxime. I imposed directional selection by increasing the concentration of cefotaxime in each round of evolution.

I found no evidence of a direct "look-ahead effect". Specifically, I show that mistranslation did not affect phenotypic evolution, as measured by an increase in resistance against cefotaxime relative to ancestral TEM-1. As expected, strong selection imposed by high cefotaxime concentrations lead to the evolutionary dynamics characterized by selective sweeps. Substitutions swept to high frequencies in a specific order in most replicate populations, regardless of the rate of mistranslation. In contrast to wild-type populations, mistranslating populations also experienced fewer fixed SNPs. However, mistranslating populations also had higher within-population genetic diversities at the end of the experiment. I show that mistranslation drove the accumulation of cryptic genetic diversity, which enhanced survival under low and intermediate concentrations of other β-lactam antibiotics.

3.2 Results

3.2.1 Experimental evolution and adaptation to cefotaxime

My study focuses on the antibiotic resistance enzyme TEM-1 β-lactamase. Wild-type TEM-1 is highly active against penicillins, such as ampicillin. To study how mistranslation may influence the evolution of a new biochemical activity, I took advantage of the fact that TEM-1 can evolve to high activity against second and third generation cephalosporins, such as cefotaxime [198–200]. Briefly, I introduced random mutations into the coding sequence of TEM-1 using error-prone PCR. I selected populations of mutated TEM-1 alleles that enabled growth on increasing concentrations of cefotaxime to the wild-type and the mistranslating hosts (figure 3.1).

Specifically, I evolved four independent populations of approximately 10^5 TEM-1 variants per host (mistranslating or wild-type), using four rounds of PCR mutagenesis and selection. In each round, I mutagenized pools of TEM-1, and cloned these pools into the original vector. I then transformed vector pools into fresh competent *E. coli* cells, divided transformed populations into subpopulations, and inoculated these subpopulations into media with increasing concentrations of cefotaxime. After 24h, I isolated plasmids from the subpopulation that survived at the the highest concentration of cefotaxime, and used this subpopulation to start the next round. Over the course of four rounds of evolution, this procedure resulted in an up to 2048-fold increase in the minimal inhibitory concentration (MIC) for cefotaxime (figure 3.2). Mistranslation reduced this MIC for ancestral TEM-1, and it did the same during all four rounds of evolution (figure 3.2A). However, relative to the ancestral TEM-1, the increase in resistance against cefotaxime was not affected by mistranslation (figure 3.2B).

I used single molecule real time sequencing (SMRT) [154] to sequence more than 500 evolved variants per population (see supplementary information, table 3.2), from each of the four rounds of evolution. Sequencing of mutated but not selected populations revealed that my mutagenesis procedure resulted in ≈ 0.7 mutations per variant per round. The mutagenesis procedure was biased towards A→G and T→C substitutions, similar to what I observed in my previous study (chapter 2).

3.2.2 Adaptation to cefotaxime is characterized by selective sweeps

My next analysis focused on the rate of sequence evolution and divergence in TEM-1 populations adapting to cefotaxime. To this end, I analyzed the sequence data from evolved populations.

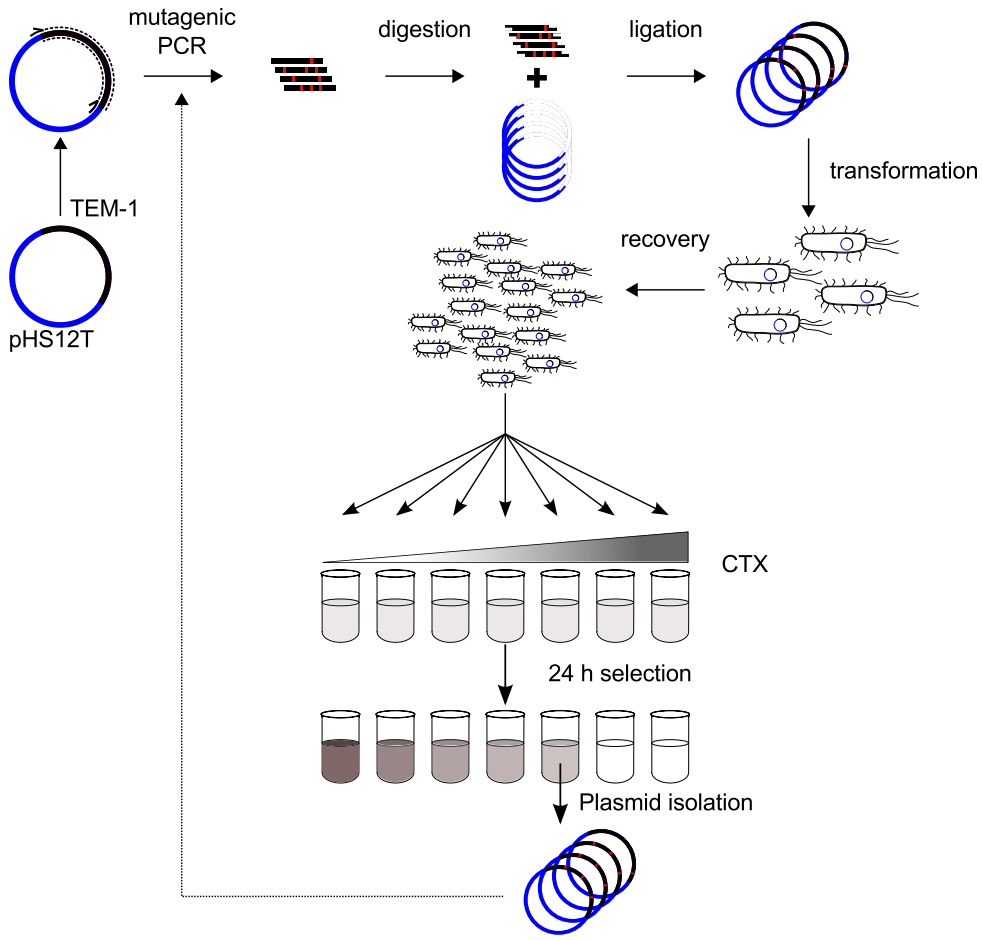

Figure 3.1: Experimental evolution of TEM-1. In each round of evolution, I exposed TEM-1 to mutagenic PCR and cloned the resulting mutant alleles into the ancestral plasmid backbone, thus ensuring that the coding sequence of TEM-1 is the only component of the experimental system that evolves. I transformed populations of plasmids with mutagenized TEM-1 into host *E. coli* cells (wild-type or error-prone). I recovered transformed cells, and divided the population into subpopulations of $\approx 10^5$ variants. I transfered each of the subpopulations into a series of liquid LB medium with a different concentration of cefotaxime, where consecutive media in the series differed by a factor 2, and selected for 24 h. Subsequently, I isolated plasmids from the subpopulation that survived at the the highest concentration of cefotaxime, and used them as templates for the next round of evolution. I evolved 4 replicate populations per host. After 4 cycles of evolution, I subjected evolved TEM-1 populations (all time points and all replicates) to single-molecule real-time (SMRT) sequencing.

I first examined substitutions that swept through the populations, i.e. SNPs that had reached frequencies above 90% at any point during my experiment. The exponential increase in resistance to cefotaxime was accompanied by selective sweeps (figure 3.3) in both hosts. The number of fixed SNPs at the end of the experiment was higher in all wild-type populations (4-8) compared to mistranslating populations (2-3).

I then examined whether mistranslation affects the frequency and the order of appearance for mutations known to be involved in adaptation to cefotaxime. Specifically, I looked for the presence of G238S, E104K, H153R, M182T, T265M, and S268G [136, 163, 201]. Of these changes, G238S and E104K function synergistically to improve the hydrolysis of cefotaxime [152], but destabilize TEM-1. Other mutations restore the stability of the enzyme, and further increase resistance. For three of these substitutions, the order of appearance and fixation of these changes was similar between populations, regardless of the host (figure 3.3C). The major evolutionary pathway G238G $\rightarrow$ E104K $\rightarrow$ M182T was followed by all populations but one (WT 3). However, there were some differences in the presence and end-point frequencies of other SNPs.

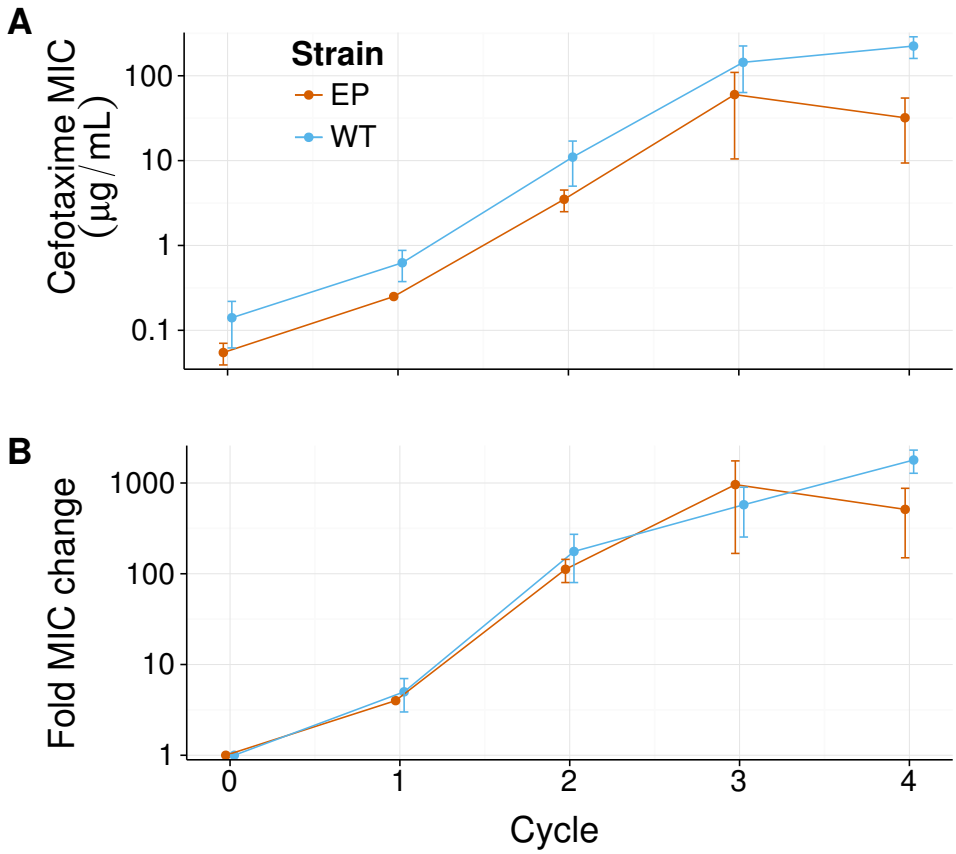

Figure 3.2: Phenotypic evolution during the experiment. (A) Mean increase in the minimum inhibitory concentration (MIC) of cefotaxime for the four replicate populations in each of the two hosts. (B) Increase in MIC values relative to the ancestral population. Error bars correspond to standard deviations across the four replicates.

The most striking difference was the absence of M182T in one of the wild type populations (WT 3). Instead, after the final round of evolution, I found H153R at a frequency of 97% in this population. In the wild-type population 1, I found T265M, which was absent from other wild-type populations, and was only present in lower frequencies in error-prone populations. Similar to WT 3, I found H153R at frequency $\approx$ 26% in the error-prone population 3, the only mistranslating population where M182T did not reach fixation. The same error-prone population has S268G at $\approx$ 51%, which was present in frequencies below 3% in all other populations.

I also examined other SNPs reaching frequencies above 90% in wild-type populations (tables 3.4 and 3.3). All of the fixed nonsynonymous SNPs other than G238S, E104K, and M182T were in fact fixed in only one of the wild type populations. In addition to nonsynonymous changes, I found five synonymous SNPs that reach frequencies above 90% (table 3.3). However, I found no synonymous SNPs that were fixed in more than one of replicate populations.

In contrast to my previous study (chapter 2), most SNPs reaching intermediate and high frequencies appeared in the structural part of TEM-1, and not in the signal peptide (tables 3.4 and 3.3). The only two fixed SNPs in the signaling peptide were I13T, and the synonymous change CTT21CTG. Both of these substitutions were fixed only in wild-type populations.

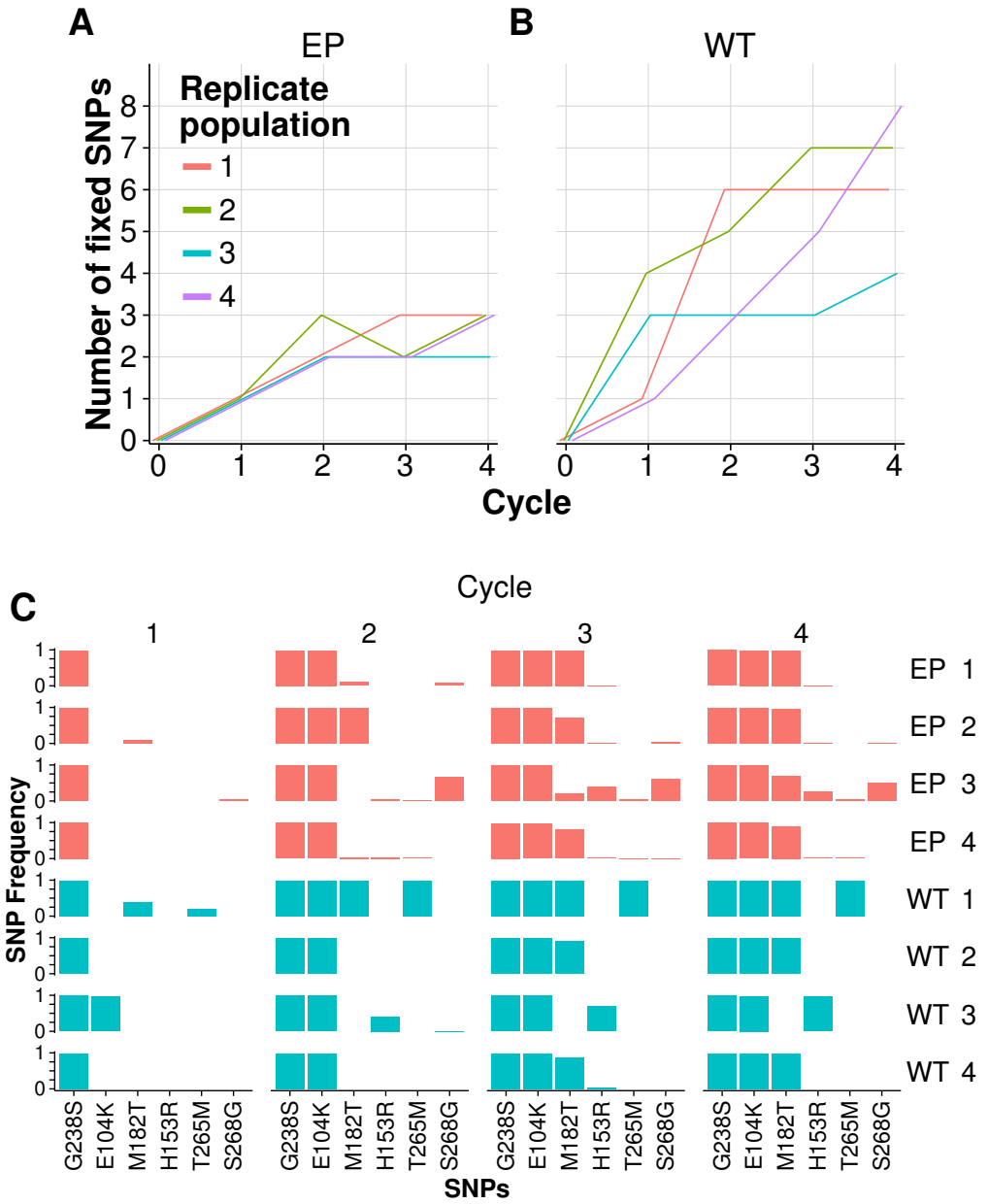

Figure 3.3: Number of fixed SNPs and SNPs implicated in resistance to cefotaxime. (A) The number of fixed SNPs (frequency greater than 90%) in error-prone populations. (B) The number of fixed SNPs in wild-type populations. (C) Frequency of SNPs known to be important for the evolution of cefotaxime resistance.

3.2.3 Mistranslation slows the rate of divergence from the ancestral TEM-1, but increases diversity within populations

In my previous study I showed that mistranslating populations evolve more slowly under purifying selection (chapter 2). I wanted to see if mistranslation affects the rate of evolution in the same way under directional selection. To this end, I calculated mean Hamming distances of evolved TEM-1 variants to the ancestral reference sequence for each of the populations. As expected for evolution under directional selection, mean distances were increasing with evolutionary time in my experiment (figure 3.4A). Mean distances were higher for all wild-type replicates in the first two rounds of evolution, and remained higher in two of the four wild-type populations at the end of the experiment, compared to mistranslating population (figure 3.4A).

In contrast to the mean distances, the standard deviations were higher in mistranslating

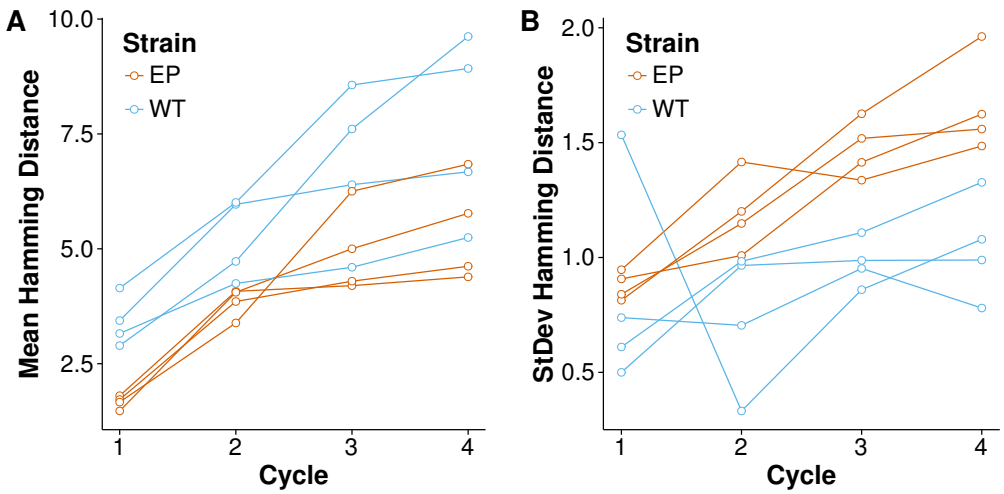

Figure 3.4: Divergence from the ancestral TEM-1. I calculated the Hamming distance for each evolved TEM-1 variant and calculated summary statistics for each population. (A) The evolution of mean Hamming distances from the reference TEM-1 for each population. (B) Standard deviation of Hamming distances for each population.

than in wild-type population (figure 3.4B).

To better characterize the spread of populations through sequence space, I computed all pairwise distances for all TEM-1 variants in each of the populations. Distributions of pairwise distances are wider and have higher means in mistranslating populations, especially in the third and the fourth round of evolution (figure 3.5).

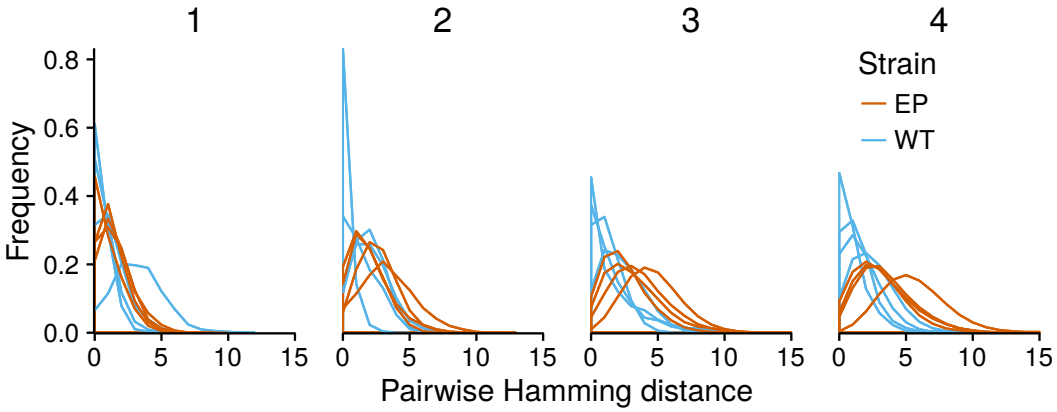

Figure 3.5: The distribution of pairwise distances in populations. I calculated all pairwise Hamming distances within each of the populations in all four rounds of evolution.

Figure 3.4B and 3.5 hint at an increase in genetic diversity in mistranslating, compared to wild-type populations. I wanted to characterize diversity further, both in nucleotide and protein sequences. To this end, I first calculated nucleotide diversities in evolved populations using the PAPNC method [202], which determines the per-population nucleotide diversity at each position of TEM-1 coding sequence from a multiple sequence alignment. Figure 3.6 (top) shows the evolution of average diversities during the experiment. In the third and the fourth round of evolution, all four mistranslating populations have a higher mean diversity than any of the wild-type populations. Furthermore, the higher diversity in mistranslating populations results from many polymorphic sites (figure 3.6 bottom) along the coding sequence of TEM-1.

Finally, I examined the diversity of individual TEM-1 variants (haplotypes) present in my

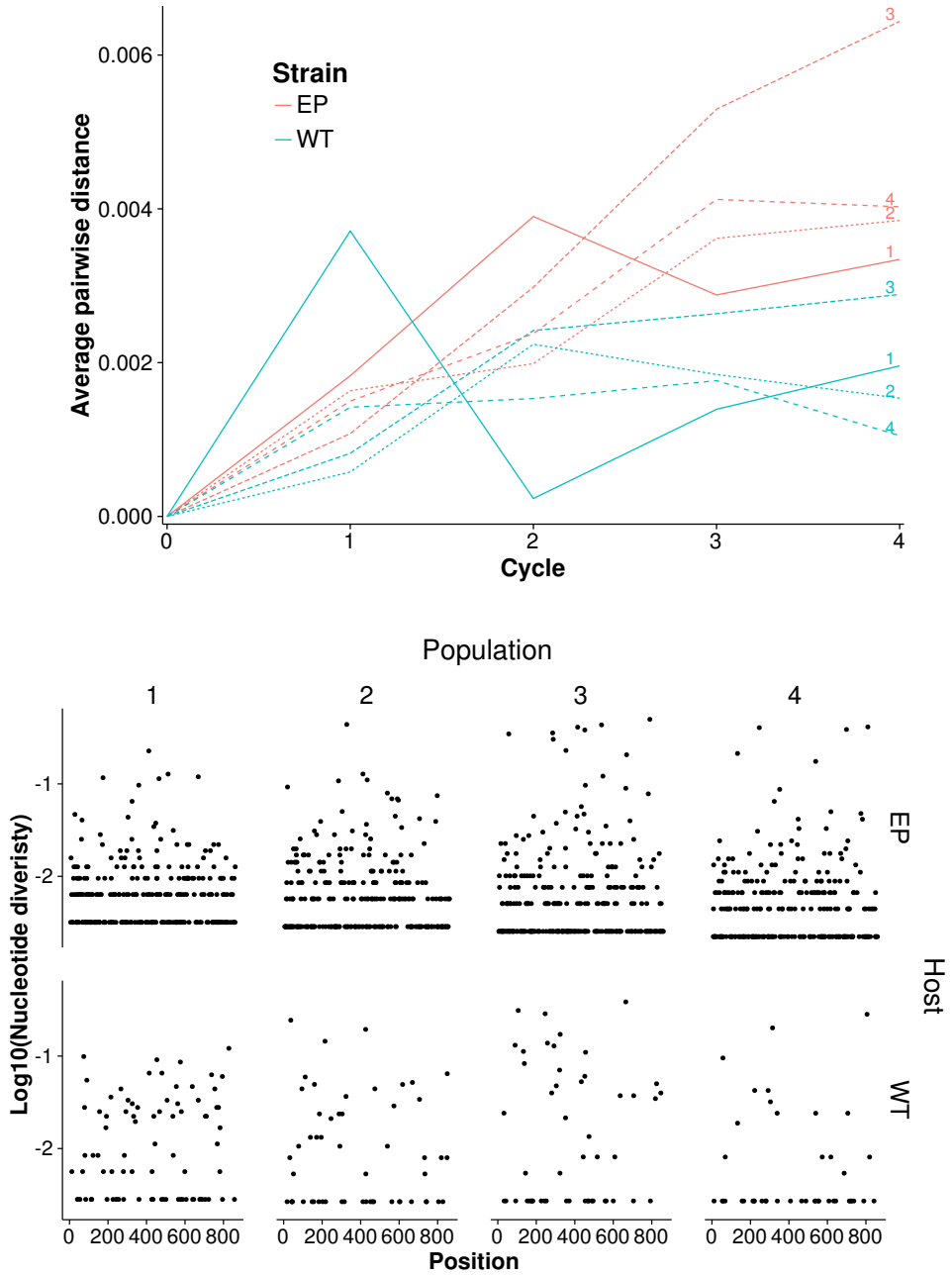

Figure 3.6: Genetic diversity in evolved populations. Top: Average pairwise diversity within evolved populations. Each line corresponds to the diversity trajectory of one of the evolving populations. Each trajectory is shown with a different line type, and the number indicates the replicate population. Bottom: Comparison of nucleotide diversities along the coding sequence of TEM-1 between error-prone and wild type populations from the final (fourth) round of evolution. I used the pairwise alignment positional nucleotide counting (PAPNC) to calculate the average diversity at each TEM-1 nucleotide site based on all sequences from each of the evolved populations. Monomorphic positions (with no diversity) are omitted from the plot.

populations. To this end, I ranked variants in each population according to their frequency, scaling the rank to the range between 0 and 1 so that I could directly compare it across different populations. I then plotted the cumulative frequency distribution of variants against their scaled rank. Wild-type populations harbor many variants with very low frequency and few variants at high frequencies (figure 3.7). In contrast, mistranslating populations contain many variants at intermediate frequencies. This variation is mostly not due to silent (i.e. synonymous) SNPs

since I observed the same pattern when I plotted cumulative frequencies of protein sequences (figure 3.7).

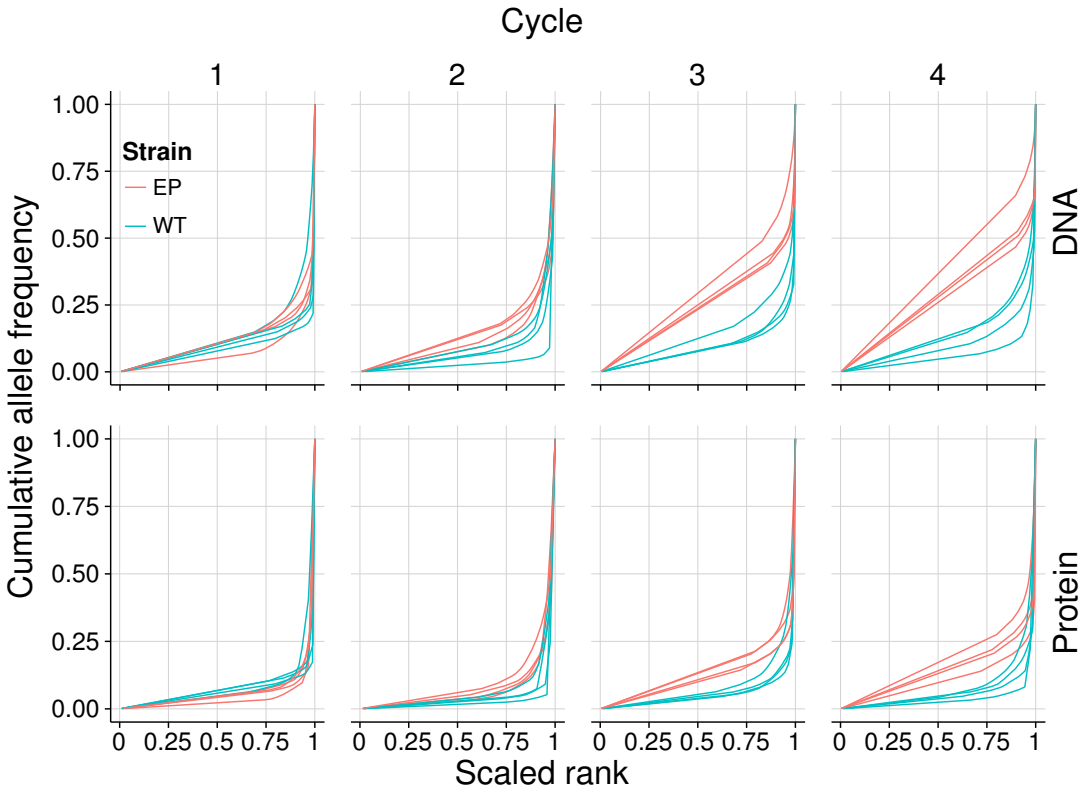

Figure 3.7: Cumulative variant (haplotype) frequencies in experimental populations. I calculated the frequency of each variant, on the DNA (top) and the protein (bottom) level, found in each of the populations from all four round of evolution. I ranked variants based on their frequency, scaled their rank to a $[0, 1]$ range, and calculated the cumulative frequency distribution for each of the populations.

To find out how different these variants are, I randomly sampled 200 sequences without replacement from each of the populations from the final round of evolution, and pooled these samples based on the host of origin. I then hierarchically clustered these sequences based on their pairwise distances. Sequences from the four wild-type populations clustered in four groups of similar sizes, while sequences from mistranslating populations could not be separated neatly (figure 3.8). This indicates a higher divergence between replicates in wild-type populations (replicates corresponding to clusters), and higher repeatability of evolution in mistranslating host.

Since sequence clustering suggests that directional evolution is more repeatable under mistranslation, I wanted to find out how distant in sequence space are populations from the fourth round of evolution. To this end, I used minimal entropy decomposition [203] on all sequences from the final round of evolution, and performed multidimensional scaling on decomposed populations. Error-prone clustered together, while wild-type populations are more distant from each other (figure 3.9).

3.2.4 Error-prone populations show increased survival under intermediate concentrations of antibiotics

I wanted to find out if some of the excess diversity from mistranslating populations (figures 3.6 and 3.8) includes TEM-1 variants that are active against different β-lactam antibiotics. To this end, I transformed TEM-1 populations from the final round of evolution into fresh mistranslating and wild-type hosts. I then tested the ability of these populations to grow in the presence of increasing concentrations of cefotaxime, ceftazidime, piperacillin, cefoxitin, clavulanic acid in

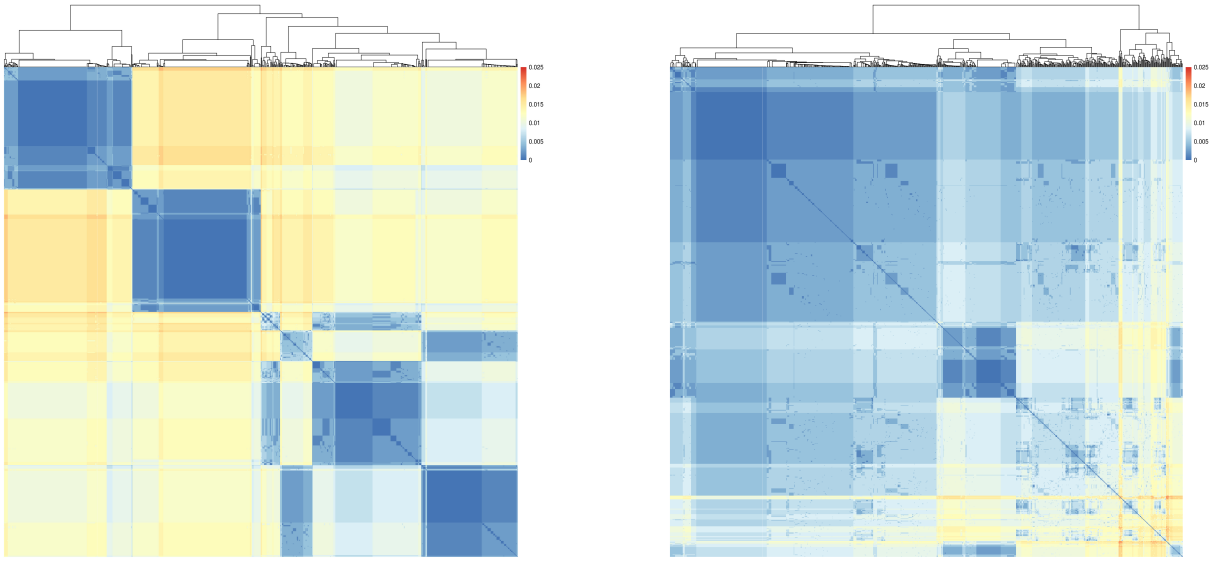

Figure 3.8: Diversity in pooled subsamples of nucleotide sequences from the final round of evolution. I randomly sampled 200 sequences from each of the populations, and then pooled them according to their hosts. I hierarchically clustered these sequences based on the nucleotide sequence identity, and created heatmaps of distance matrices. Dark blue color corresponds to blocks of similar sequences, yellow and red show higher divergence. The heatmap is symmetric, i.e. the second dimension carries no additional information.

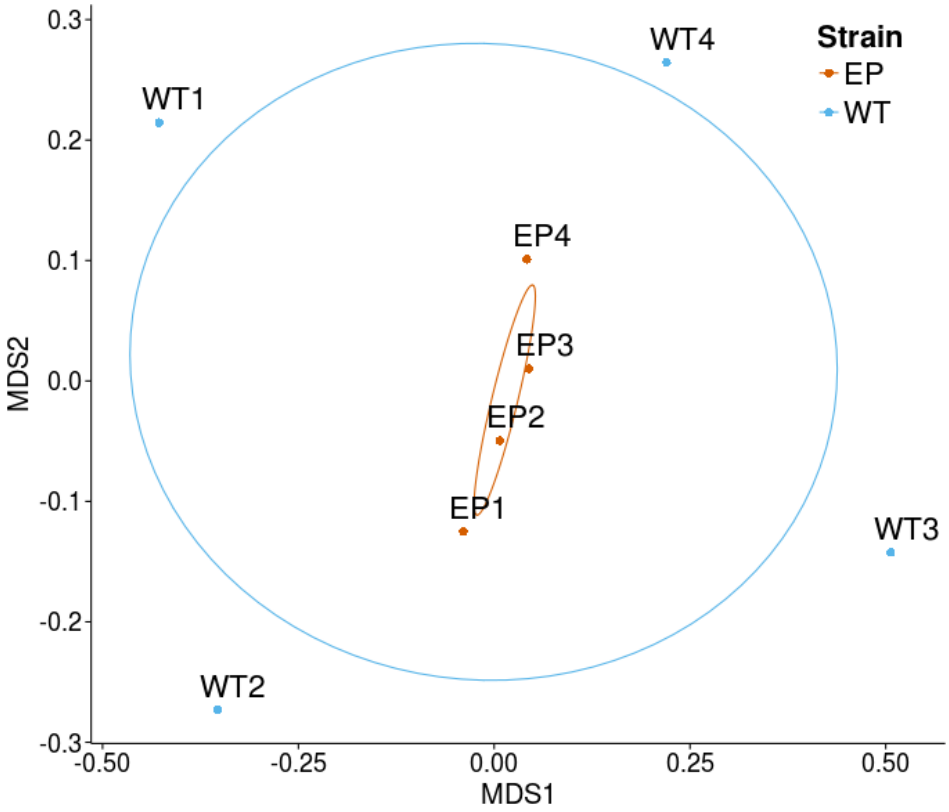

Figure 3.9: Multidimensional scaling and clustering of populations from the final round of evolution. I decomposed nucleotide sequences from the fourth round of evolution using minimal entropy decomposition (MED) and performed multidimensional scaling on decomposed populations using the Canberra distance measure.

combination with cefotaxime, and oxacillin in combination with clavulanic acid. Populations evolved in mistranslating hosts enabled growth to higher mean population densities under low and intermediate concentrations of all tested antibiotics (figure 3.10). This was true whether TEM-1 variants evolved under mistranslation were expressed in wild-type or mistranslating hosts.

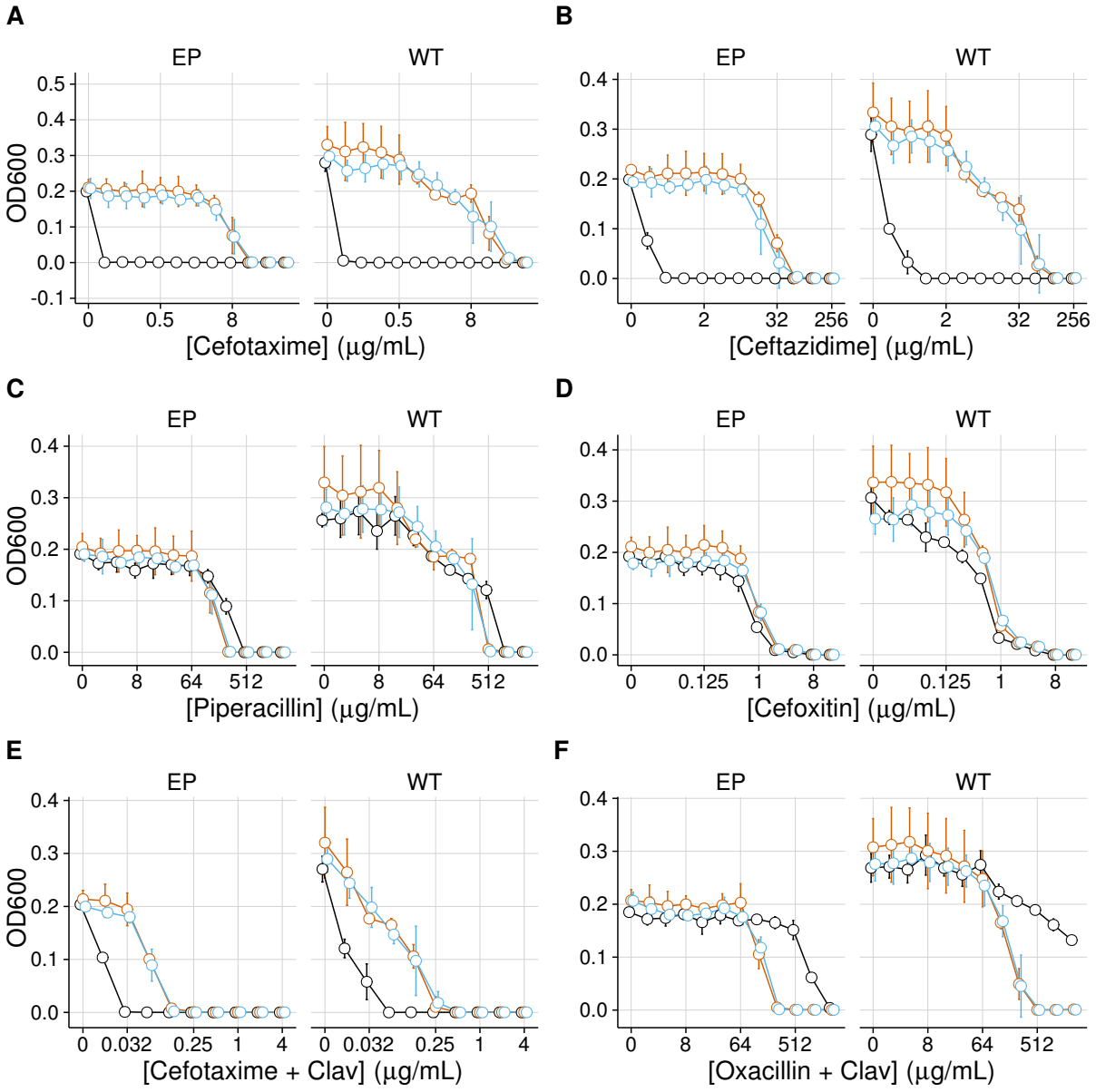

Figure 3.10: Optical densities (OD_{600}) of evolved populations measured in media with different β-lactams and β-lactamase inhibitors (Clav = clavulanic acid). Each population evolved in wild-type hosts (blue) and error-prone hosts (red) was expressed in both wild-type (WT) and error-prone (EP) hosts. I used ancestral TEM-1 as a control in these experiments (shown in black). Transformed cells were allowed to recover and then exposed to LB media with different concentrations of β-lactam antibiotics. Optical density was measured at 600 nm after $\approx$ 22 hours. Optical density of population was computed as the mean from at least four independent experiments, and circles correspond to means across all four populations. Error bars are standard deviations across four populations.

3.3 Discussion

I sought to study the effect of mistranslation on the evolution of new biochemical activities in TEM-1 β-lactamase. More specifically, I used directed evolution of TEM-1 β-lactamase under elevated mistranslation rates to test the "look-ahead effect" hypothesis [135]. The evolution of resistance to cefotaxime in TEM-1 is a good candidate system to study the "look-ahead effect" because it requires multiple substitutions. Furthermore, the order of appearance of these substitutions is constrained [136]. Phenotypic mutations could in principle accelerate the evolution of resistance to cefotaxime. These mutations could also change the accessibility of mutational paths. However, I found no evidence of "the look-ahead effect" under directional selection for resistance against cefotaxime.

First, there was no major difference in the evolution of resistance against cefotaxime between the error-prone and wild type hosts (figure 3.2). While the relative increase in cefotaxime resistance was similar in both hosts, the absolute resistance, measured by MIC, was even lower in mistranslating populations. Second, error-prone populations followed similar evolutionary trajectories as wild-type populations. Specifically, genotypic evolution proceeded by successive selective sweeps of three main SNPs, G238S, E104K, and M182T (figure 3.3 C). Third, wild-type populations harbored more SNPs at high frequencies than mistranslating populations (figure 3.3 A and B), suggesting that mistranslation constrains adaptation to cefotaxime, rather than facilitating it. Finally, replicate mistranslating populations were more similar in their genetic composition at the end of the experiment.

In contrast to the lack of evidence for a direct "look-ahead effect", I found that mistranslation increased within-population genetic diversity under directional selection. Both DNA and protein diversity were greater in error-prone compared to wild-type populations at the end of the experiment. Furthermore, I show that this cryptic genetic diversity improves survival under low and intermediated concentrations of β-lactams other than cefotaxime, in both mistranslating and wild-type hosts (figure 3.10).

3.3.1 Phenotypic evolution

Mistranslation can have deleterious effects because it destabilizes proteins. This reduces the concentration of active TEM-1 and can cause other pleiotropic effects that increase sensitivity to some antibiotics. Since the resistance to β-lactams is dependent on the combined effects of the activity and the concentration of TEM-1, it is not surprising that the mistranslating host with ancestral TEM-1 has a lower absolute MIC for cefotaxime, compared to the wild-type host (figure 3.2A). Similarly, throughout the evolution experiment, the absolute MIC values remain lower in mistranslating population. However, the relative increase in MIC follows the same trajectory in both hosts (figure 3.2B). This suggests that mistranslation has no effect on the rate, or the end point of phenotypic evolution in my experiment.

Theory suggests that phenotypic plasticity and molecular noise can facilitate adaptation, as long as they are increased in the direction of selection [1, 204]. During adaptation of TEM-1 to cefotaxime, this can happen when cefotaxime-adaptive mutations, such as G238S, E104K, and M182T, appear through mistranslation. However, even at the highest bacterial mistranslation rates, the active enzyme concentration might insufficient to survive cefotaxime selection, because synthesis of only a very small fraction of protein variants with increased activity against cefotaxime is possible. Because the resistance to β-lactam antibiotics is determined by the concentration of β-lactamase, the beneficial effect of mistranslation is nullified by the high concentration of cefotaxime in the media. In other words, cells can only survive selection if they genetically encode beneficial alleles because they produce higher concentrations of the active enzyme.

3.3.2 Selective sweeps

Previous studies on the evolution of TEM-1 activity on cefotaxime and other extended spectrum cephalosporins revealed five substitutions that often occur in combination in laboratory and clinical isolates [136, 198, 200]. These changes include one noncoding (g4205a), and four coding substitutions (A42G, E104K, M182T, and G238G). In my experiment only the latter could occur, because I constrained the mutagenesis to the coding part of the TEM-1. Of the cefotaxime-adaptive changes, I observed E104K, M182T, and G238G. A42G did not appear in any of the variants I sequenced. There is some evidence that A42G is stabilizes the active site of TEM-1 [200]. Thus, the absence of A42G in my and other similar evolution experiments [201] can be compensated for by other stabilizing substitutions [201]. Indeed, I found T265M and other stabilizing substitutions present at high frequencies in evolved populations (table 3.4).

The order of appearance and fixation for the three main mutations is typically conserved (G238G $\rightarrow$ E104K $\rightarrow$ M182T) [136, 201, 205]. G238G and E104K jointly improve cefotaxime binding and hydrolysis, but have destabilizing effects [206]. The third substitution, M182T, is a global supressor [206]. That is, it suppresses destabilizing effects of other mutations and thus further increases the resistance to cefotaxime. M182T is frequent in clinical isolates [163, 206], marking the importance of M182T in the evolution of TEM-1. In my experiment, the exponential increase in cefotaxime MIC was accompanied by selective sweeps of G238G, E104K and M182T in almost all replicate populations (figure 3.3 A-C), regardless of the rate of mistranslation. M182T was absent from only one of the wild-type populations (figure 3.3C, WT 3), but its stabilizing effect could have been compensated by H153R [207] and A224V [161] (table 3.4) . Similarly, I found one error-prone population where M182T appeared, but never reached a frequency above 90% (figure 3.3C, EP 3). This population harbored H153R, as well as S286G, which can also stabilize TEM-1 [163].

Curiously, out of all other stabilizing SNPs that become fixed in my wild-type populations (I47V, N100D, I208M, T265M, and K288E), I47V is the only one found at high frequency in one of the mistranslating populations ($\approx$ 88%). Other SNPs fixed in wild-type populations are private, i.e. found at high frequencies in only one population (table 3.4). Furthermore, N100D, I208M, T265M, and K288E, all become fixed in their respective wild-type populations as early as in the first or the second round of evolution (table 3.4). This is contrary to the expectation that selection should strongly favor fixation of stabilizing SNPs in error-prone populations. The finding that stabilizing SNPs are preferentially found in wild-type populations suggests that selection for resistance to cefotaxime is much stronger than selection for stability in mistranslating hosts. The appearance of G238G destabilizes TEM-1, and slows down the accumulation of other nonsynonymous SNPs in mistranslating conditions, because nonsynonymous SNPs tend to destabilize proteins [168]. Even when potentially stabilizing SNPs occur in the population, their fixation is prevented by strong selection that drives alleles with G238G (in the first round of evolution) and E104K (in the second) to high frequency.

Another possible explanation for the fixation of stabilizing SNPs in wild-type populations is the absolute concentration of cefotaxime used in selection. Absolute MIC values are consistently higher for wild-type populations throughout the experiment (figure 3.2A and table 3.1). Additional stabilizing SNPs might be necessary for alleles in wild-type population to survive cefotaxime concentrations that are typically 2-4 times higher than for mistranslating populations. The lack of parallelism, i.e. that these SNPs fix only in one replicate population, supports the claim that these SNPs have weak beneficial effects at best.

Strong selection for activity against cefotaxime, rather than for increased stability or expression, is supported by further two findings. First, none of the synonymous SNPs reach high frequency in mistranslating populations. Second, of all fixed SNPs, only one synonymous and one nonsynonymous occurred in the signaling peptide. The signal peptide of TEM-1 regulates the level of expression and periplasmic concentration, and scarcity of changes in this region reflects weak selection for increased or decreased expression of TEM-1.

3.3.3 Mistranslation creates cryptic genetic variation under directional selection

When an environment changes, a population can find itself poorly adapted to the new environment. This can lead to a mode of selection called directional selection. Whether directional selectional causes an increase or decrease in genetic diversity is still unclear, and may depend on many factors [208]. Previous study found that directional selection can increase genotypic diversity in laboratory populations of evolving ribozymes, while maintaining phenotypic diversity [208].

I find that populations of TEM-1 alleles evolving under directional selection and increased mistranslation rates increase in genetic diversity (figures 3.5). This excess genetic diversity is not synonymous, i.e. it gives rise to diverse protein products (figure 3.7). A possible explanation lies in less efficient selection in mistranslating populations. Pleiotropic effects of mistranslation on cellular physiology increase the sensitivity to cefotaxime. This is demonstrated by the lower MIC of ancestral TEM-1 in mistranslating strain (table 3.1), and lower absolute cefotaxime MIC values during the evolution (figure 3.2A and table 3.1). Thus, mistranslating populations are experiencing selection at lower concentrations of cefotaxime than wild type populations. This reduction in the strength of selection promotes clonal competition, and enables stable coexistence of many SNPs and alleles at intermediate frequencies (figures 3.6 and 3.7).

Reduced effective population size provides another explanation for the reduction in the strength of selection under elevated mistranslation. Stochastic gene expression noise can create a permanent genetic load, which in turn reduces effective population size and enhances the effect of drift on the population level [209]. Very similar to stochastic gene expression, mistranslation increases the genetic load by producing cytotoxic misfolded proteins [22], and this can cause the reduction of the effective population size, weakening selection.

Weaker selection enables the maintenance of genetic variation in a population even if it is neutral with respect to cefotaxime resistance. This genetic diversity can promote the evolvability of a population [210]. Upon an environmental change, more diverse populations can ensure that some alleles are already pre-adapted to the new environment [211]. Indeed, I find that populations evolved under mistranslation enable growth to higher population densities when challenged with low to intermediate concentrations of novel β-lactam antibiotics (figure 3.10).

High concentrations of β-lactams impose strong selection on a population, and kill off all but the fittest individuals (alleles). In other words, the MIC of a population will depend on the activity of the fittest allele in the population, and this upper limit might be the same for populations evolved in mistranslating and wild-type hosts. However, under low and intermediate concentrations of antibiotics, other traits, such as stability and expression costs, can be important for achieving higher population density. Mistranslating populations are more likely to harbor alleles with those traits, since their diversity is higher than the diversity of wild-type populations.

3.3.4 Mistranslation increases the repeatability of evolution

The ”look-ahead effect” depends on the phenotypic diversity created by mistranslation to facilitate adaptive evolution. The astronomically large number of protein variants that can be created through phenotypic mutations suggests that adaptation under mistranslation might be a highly stochastic process, driving populations through many different mutational pathways and to different end points. In contrast, in my experiments elevated mistranslation gives rise to replicate populations that are genetically closer than wild-type populations (figures 3.8 and 3.9). One possible explanation is that the evolution of resistance to cefotaxime is highly constrained and can follow only a few paths [136]. Stability constraints caused by mistranslation might further reduce the number of accessible paths. However, this explanation is not likely because the sequence landscape of TEM-1 contains many variants with high activity against cefotaxime [194]. These variants require multiple simultaneous substitutions, but should in principle be accessible through mistranslation, allowing for multiple evolutionary

end-points.

It is more likely that repeatability under elevated mistranslation is caused by population-genetic constraints. The reduction in selection coefficients of beneficial substitutions in mistranslating populations prevents selective sweeps, and constrains the diffusion away from the reference TEM-1. Thus, at the end of the experiment, mistranslating populations have diverged less from each other, and cover overlapping regions of the sequence space.

My observations highlight the importance of studying phenotypic mutations and their effect on protein evolution. Taken together with observations from the previous chapter, my findings show that mistranslation can affect the evolution of protein stability, gene expression, and cryptic genetic variation.

3.4 Materials and methods

3.4.1 Media and antibiotics

I used Difco LB broth (BD) for all experimental steps involving growth and selection. For preparing competent cells, I used SOB media (Sigma). For recovery of electroporated cells I used SOC media (SOB media with 20 mM glucose). For antibiotic selection, I used chloramphenicol at 25 and 34 μg/l, and cefotaxime sodium salt (Sigma) at 25, 100, and 250 μg/L. I used saline (0.9 μg/L NaCl) to prepare serial dilutions for library size estimations.

3.4.2 Strains

I used the *E. coli* strain DH5α for all cloning steps, including the preparation of pre-selection TEM-1 libraries. The construction of the ribosomal mutant with increased mistranslation rate, *rpsD12*, was described elsewhere [145]. The original *rpsD12* allele was kindly provided by T. Nyström [30]. I transfered this allele into a fresh genetic background using PCR-based recombineering [178]. Specifically, I first integrated a kanamycin resistance cassette flanked by flippase recognition target (FRT) sites [178] into the genome of the *rpsD12* strain. Using PCR, I amplified the region spanning the mutation and the resistance cassette. I then integrated the fragment into a MG1655 background (CGSC#7740). Next, I used P1 transduction to transfer the mutation linked to the kanamycin resistance cassettes into the fresh MG1655 background. Finally, I removed the KanR cassette from the genome with flippase plasmid pCP20 [179]. I picked clones that were sensitive to kanamycin (because the resistance cassette was excised, which leaves an FRT scar behind [178]), and confirmed the construct by Sanger sequencing. I repeated the same procedure with the wild-type MG1655 strain, leading to a control I refer to as the wild-type or normal strain in the rest of the chapter. This construct had a wild-type rpsD allele, and an FRT site at the same location as the error-prone *rpsD12* strain.

3.4.3 Plasmids

I used a high copy number plasmid with a chloramphenicol resistance marker, based on pHSG396 [174] for evolving TEM-1. To create this plasmid, I first removed the lac promoter from the pHSG396 plasmid and introduced the coding region of TEM-1 with its constitutive promoter (pAMP) from pBR322. In this plasmid, which I call pHS13T, the coding region of TEM-1 is flanked by SacI and HindIII restriction sites. To facilitate gel extraction of vector backbones for recloning, I constructed a second vector, pHS13K, which differed from pHS13T by having a KanR cassette from pKD4 [178] as a filler sequence between SacI and HindIII sites.

3.4.4 Electrocompetent cells

I prepared electrocompetent cells by glycerol/mannitol density step centrifugation [180]. In short, I diluted 2 mL of an overnight culture in 200 mL SOB media. I incubated this culture with shaking (250 rpm) in a 2L shake flask at 37°C, until the OD_{600} reached values of 0.4-0.6. I chilled the culture in iced water for 15 min, and centrifuged at 1500 g and 4°C, for 15 min in a 5810-R Eppendorf centrifuge with the F-34-6-38 rotor. I resuspended the pellet in 40 mL of ice-cold ddH_2O, and split it into two 50 mL tubes. I then slowly added 10 mL of ice-cold 20% (w/v) glycerol + 1.5% (w/v) mannitol to the bottom of each tube with a 12 mL pipette. I centrifuged the suspension again at 1500 g at 4°C for 15 min, with acceleration/deceleration set to zero. I removed the supernatant by aspiration, and resuspended the pellet in 1 mL of ice-cold 20% (w/v) glycerol + 1.5 % (w/v) mannitol. I aliquoted the cell suspension (80 μL for DH5α, and 100 μL for *rpsD12* and wt strains) in 1.5 mL tubes, froze them in a dry ice-ethanol bath, and stored at -80°C.

3.4.5 Mutagenesis

To introduce genetic diversity into TEM-1 populations, I used a 25 cycle error-prone PCR with nucleoside analogues [155]. A 100 μL PCR reaction contained 10 ng of the template plasmid (pHS13T in the first round, and selected plasmid population in subsequent rounds), with 400 μM dNTPs (Thermo Scientific), 2.5 U Taq polymerase (NEB), Thermopol buffer (NEB), 3 μM 8-oxo-GTP, 3 μM dPTP (Trilink Biotechnologies), as well as 400 nM of primers TEM1-F6 and TEM-R6 (table 3.5). To remove the template plasmid, I treated the PCR product with the restriction enzyme DpnI for 2 h at 37°C. Subsequently, I inactivated all enzymes by adding 0.6 U of proteinase K (Thermo Scientific) and incubated for 1h at 50°C, followed by a 15 min proteinase K inactivation at 80°C.

3.4.6 Library cloning

I carried out the double restriction of the mutagenized TEM-1 pool with 20 U of SacI-HF and HindIII-HF (NEB) for 2h at 37°C, followed by 20 minutes at 80°C. I then purified double digested inserts with the QIAprep PCR purification kit (Qiagen) and eluted them in 2.5 mM Tris-Cl, pH 8.5. In parallel, I double digested the plasmid backbone by incubating pHS13K with 20 U of SacI-HF and HindIII-HF for 16 h. I gel purified the digested vector and dephosphorylated it by incubating with 5 U of Antarctic Phosphatase (NEB) for 1 h, followed by a 20 minute inactivation at 80°C. I then ligated 19 ng of insert (TEM-1 pool), and 50 ng of digested and dephosphorylated vector in 20 μL reactions with 10 U of T4 DNA ligase (NEB) for 16h at 4°C. I inactivated the T4 DNA ligase by incubating for 10 min at 65°C. I precipitated the ligation product by adding 80μL of H_2O, 20 μg of glycogen (Thermo Scientific), 50 μL of 7.5 M ammonium acetate (Sigma), and 2.5 volumes of ice-cold ethanol. I incubated the mixture at -80°C for 20 min, centrifuged for 20 min at 18000 g, washed in 800 uL of 70% cold ethanol, centrifuged and washed again. I dried the pellet under vacuum for 15 min, and then resuspended in 15 uL of 2.5 mM Tris-Cl, pH 8.5.

3.4.7 Preselection libraries

Because I derived the wild-type and the *rpsD12* strain from a restriction-positive MG1655 strain, direct transformation of non-methylated ligation products would result in low transformation efficiency due to restriction. To ensure plasmid methylation before selection in wild-type and *rpsD12* strains, I transformed ligation products into restriction-deficient DH5α cells. To this end, I mixed 80 μL of electrocompetent DH5α cells with 4 μL of the precipitated ligation product, and electroporated using a Micropulser electroporator (Bio-Rad) set on EC3 (15 kV/cm), and 0.2 cm electroporation cuvettes (Cell Projects). Immediately after electroporation, I added 1 mL of pre-warmed SOC media to transformed cells, and transfered the suspension to a 24-well

plate. I allowed cells to recover by incubating the plate at 37°C with shaking at 400 rpm for 1.5 h. After the recovery period, I centrifuged the plate and aspirated the supernatant from the plate. I resuspended the cell pellet in 5 mL of LB media supplemented with 34 ug/mL of chloramphenicol. I used a 50 μL cell suspension aliquot to estimate library size by making serial dilutions in saline, and plating on LB agar with 20 μg/mL chloramphenicol. Through this procedure, I estimated library sizes to lie between 10^5-10^6 sequences. I incubated transformed cells overnight at 37°C with shaking at 320 rpm. The next morning, I stored 1 ml of the overnight culture as a glycerol stock, and used the rest to purify plasmids with a QIAPrep miniprep kit (Qiagen).

3.4.8 Selection

I transformed 100 μL aliquots of electrocompetent *rpsD12* or wt cells with approximately 5 ng of purified preselection libraries. The electroporation conditions were the same as for preselection libraries. After 1.5 h of recovery in 1 mL SOC media, I centrifuged the recovered cell suspension for 10 min at 2800 g, and resuspended cell pellets in LB media with 34 μg/mL chloramphenicol. From each of the resuspended libraries, I inoculated approximately 10^5 cells into a two-fold dilution series of cefotaxime (the highest concentration of cefotaxime used in the experiment was 2048, and the lowest 0.0078 μg/mL) in LB media with 34 μg/mL chloramphenicol. Selection lasted for approximately 22 h with shaking at 320 rpm at 37°C. I isolated plasmids using the QIAPrep miniprep kit (Qiagen) from the highest concentration of cefotaxime where growth was visible. These plasmids were then used as a starting point for the next round of evolution.

3.4.9 Control libraries

To estimate mutation rates in each mutagenesis cycle, I constructed one control library for each host strain. These libraries were subject to the same procedure as libraries under selection, except that the selection media contained only 34 ug/mL chloramphenicol and no ampicillin. I subjected these control libraries to a single round of evolution.

3.4.10 Antibiotic susceptibility assays

I wanted to test the ability of ancestral and evolved TEM-1 populations to confer resistance to different β-lactam antibiotics in the two hosts. To this end, I transformed electrocompetent wild-type and mistranslating strains with pHS13T plasmid carrying ancestral TEM-1. I also transformed populations evolved in wild-type hosts into both wild-type and error-prone hosts, and did the same with populations evolved in error-prone hosts. After the recovery period (1.5 h in SOC media, at 37 °C, shaking at 400 rpm), I centrifuged cultures for 10 min at 2200 g, aspirated the supernatant, and resuspended the cell pellet in 4.5 mL of LB supplemented with 34 μg/mL chloramphenicol. I grew these cultures for 5 hours at 37 °C with shaking at 320 rpm, and then stored them as glycerol stocks at -80°C.

On the morning of the susceptibility assay, I scraped frozen cultures, and inoculated them in 96-deep well plates (Nunc) with 1.1 mL of LB and 34 μg/mL chloramphenicol. I grew these cultures for ≈5 hours, measured their OD_{600} and diluted them to an OD_{600} of 0.01. I inoculated 10 μL of the diluted culture into 96-well plates with 190 μL of LB supplemented with chloramphenicol, and 2-fold dilutions of cefotaxime, ceftazidime, cefoxitin, piperacillin, cefotaxime + clavulanic acid (0.1 μg/mL), and oxacillin and clavulanic acid (0.5 μg/mL). I incubated these plates at 37 °C without shaking, and measured OD_{600} after 22 hours.

3.4.11 SMRT sequencing

I amplified libraries from all the experimental populations in a two step PCR (figure 3.11). In the first step, I used a 25-cycle PCR with the Phusion polymerase to amplify the coding

region of TEM-1 with TEM1FS-F and TEM1FS-R primers (table 3.5). I gel purified PCR products and used them as templates for a second 25-cycle PCR with barcoded primers BCXX and ELP (see table 3.5 for primer sequences), using 6 bp-long barcodes described in [181]. I purified PCR products from the second PCR using a QIAprep PCR purification kit (Qiagen). To check the quality and concentrations of amplicons in each library, I used the Agilent 2200 TapeStation System (Agilent Technologies). To account for sequencing and library preparation errors, I amplified and barcoded an additional library from an ancestral TEM-1 sequence. Finally, I combined 20 ng of DNA from each library to create a final amplicon pool for sequencing.

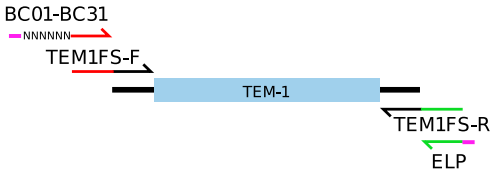

Figure 3.11: Primers used in the two-step PCR for uniquely barcoding populations.

I produced a SMRTbell library from the amplicon pool with the DNA Template Prep Kit 2.0 (250bp - 3Kb) (Pacific Biosciences p/n 001-540-726). To this end, I inspected amplicon size and integrity on an Agilent Bioanalyzer 2100 1Kb DNA Chip. I then used polishing enzymes to end-repair 500-750 ng of DNA from the amplicon pool. Subsequently, I created the SMRTbell template by blunt end adapter ligation. After that, I used the Agilent Bioanalyzer 12kb DNA Chip and a Qubit Fluorimeter (Life technologies) to confirm the quality of DNA in the library and estimate its concentration. Finally, I used a DNA/Polymerase P4 binding kit (Pacific Biosciences) according to the manufacturer instructions to create a ready-to-sequence SMRTbell-polymerase complex. I programmed the Pacific Biosciences RS2 instrument to sequence the library on two SMRT cells v3.0 (Pacific Biosciences), using P4/C2 chemistry, the magnetic bead loading method, and taking two movies of 180 minutes for each cell. After the sequencing run, I generated a sequencing report via the SMRTportal, in order to assess adapter dimer contamination, sample loading efficiency, average read-length, and the number of filtered sub-reads.

3.4.12 Primary data analysis

I assembled consensus reads of TEM-1 variants (reads of insert) from subreads using the SMRTAnalysis v2.3 package. I filtered reads of insert according to a) the minimum number of full pass subreads (4), b) the minimum predicted consensus accuracy (0.9), and c) read of insert length (850-1200 bp). With a mean number of ≈12.3 passes per read of insert, this procedure resulted in 51,034 reads, with a mean read length of 979 bp, and an average read quality of ≈0.98.

I mapped reads to the reference (ancestral) TEM-1 sequence using BLASR [175] with a minimum accuracy of 0.9, and a minimum mapped length of 850 bp. The resulting total number of mapped reads was 51,365, with the average mapped read length being 973 bp. The mean mapped subread concordance was 0.976 I further filtered mapped reads to include only those reads with average Phred quality > 20, and spanning the entire coding region of the TEM-1 reference in the alignment. I demultiplexed the filtered set of reads according to their barcodes, using custom Python scripts based on the *pbcore* module. The final set of reads contained only sequences whose barcodes perfectly matched those I used during library preparation.

Because indels are a major source of errors for SMRT sequencing, and because more than 98 % of indels in TEM-1 are loss-of-function mutations [170], I focused my analysis on point mutations. I considered a mismatch of a TEM-1 sequence read to the reference TEM-1 sequence a true SNP only if its Phred quality score was above 20 (see table 3.2 for summary statistics). I repeated all the analyses with the Phred quality score filter of 0, 10, 30 and 40, but this did not qualitatively change any of my results.

3.4.13 Genetic diversity calculations

I used the pairwise alignment positional nucleotide counting (PAPNC) method [202] to calculate the per-site and average genetic diversity in populations of TEM-1. Specifically, I first calculated the genetic diversity at site j as:

$$D_j = A_j \times (C_j + G_j + T_j) + C_j \times (G_j + T_j) + G_j \times T_j \tag{3.1}$$

where $A_j, C_j, G_j,$ and T_j are the numbers of bases A, C, G, T, respectively, at site j in the alignment. The per-site distance is expressed as:

$$PD_j = \frac{D_j}{SP} \tag{3.2}$$

where $SP = \frac{N \times (N-1)}{2}$, and N is the number of nucleotides at position j. The average per-site distance is:

$$APD = \frac{\sum_{i=1}^{L} PD_j}{L} \tag{3.3}$$

For visualizing similarity and identity in protein in nucleotide sequence alignments, I used *R* 3.1.3. I first computed identity and similarity matrices from sequence alignments using the Seqin*R* package 3.1-3 [212], and the Fitch method [213]. I used pheatmap package 1.0.7 to plot heatmaps from matrices.

3.4.14 Minimal entropy decomposition

To partition sequence data into clusters, I used minimal entropy decomposition (MED) [203]. MED is an algorithm that partitions large sequence datasets based on information-rich nucleotide positions. Specifically, I applied the function *decompose* from the Oligotyping software package [214] to the sequence data from the final round of evolution, using default parameters. To reduce the dimensionality and visualize the results of MED analysis, I used multidimensional scaling function with Canberra distance measure *metaMDS* from the *vegan* R-package version 2.3-1 [215] on the MED output matrix.

3.5 Supplementary information

3.5.1 MIC values during selection

Table 3.1: Cefotaxime MIC values in the selection step of experimental evolution. Values are in μg/mL. MIC(Ancestral) is the mean ± standard deviation of four replicates, measured for mistranslating and wild-type hosts carrying pHS13T plasmid with the ancestral TEM-1, using the same media as for experimental populations.

Strain	MIC(Ancestral)	Population	MIC(Cy1)	MIC(Cy2)	MIC(Cy3)	MIC(Cy4)
Mistranslating)	0.055 ± 0.014	1	0.25	4.00	128.00	64.00
		2	0.25	4.00	64.00	32.00
		3	0.25	2.00	16.00	16.00
		4	0.25	4.00	32.00	16.00
Wild-type)	0.141 ± 0.070	1	0.50	16.00	256.00	256.00
		2	1.00	8.00	128.00	256.00
		3	0.50	16.00	128.00	128.00
		4	0.50	4.00	64.00	256.00

3.5.2 Sequencing library statistics

Table 3.2: Sequencing and SNP statistics. Population names are given in the format Host_Replicate_Cycle. EP and WT refer to error-prone, and wild-type host, respectively. The number of SNPs refers to observed SNPs before (raw), and the number after quality filtering (HQ)

Library	Reads	Mean Quality	SNPs (raw)	SNPs (HQ)	SNPs per read (raw)	SNPs per read (HQ)
EP_L1_C1	661	40.7	1409	1197	2.13	1.81
EP_L2_C1	723	40.7	1599	1258	2.21	1.74
EP_L3_C1	691	40.8	1182	1019	1.71	1.47
EP_L4_C1	734	40.9	1532	1230	2.09	1.68
EP_L1_C2	707	40.6	3321	2883	4.70	4.08
EP_L2_C2	721	40.8	3037	2782	4.21	3.86
EP_L3_C2	823	40.6	3676	3338	4.47	4.06
EP_L4_C2	671	40.6	2618	2275	3.90	3.39
EP_L1_C3	645	40.6	3089	2708	4.79	4.20
EP_L2_C3	673	40.5	3329	2891	4.95	4.30
EP_L3_C3	584	40.5	3361	2921	5.76	5.00
EP_L4_C3	705	40.4	4961	4408	7.04	6.25
EP_L1_C4	626	40.4	3290	2748	5.26	4.39
EP_L2_C4	698	40.6	3840	3225	5.50	4.62
EP_L3_C4	785	40.4	5422	4531	6.91	5.77
EP_L4_C4	896	40.4	6814	6130	7.60	6.84
WT_L1_C1	666	40.8	2755	2305	4.14	3.46
WT_L2_C1	743	40.9	3242	3104	4.36	4.18
WT_L3_C1	750	40.9	2550	2394	3.40	3.19
WT_L4_C1	755	40.8	2437	2218	3.23	2.94
WT_L1_C2	722	40.8	4379	4308	6.07	5.97
WT_L2_C2	766	40.7	4920	4605	6.42	6.01
WT_L3_C2	856	40.6	4007	3635	4.68	4.25
WT_L4_C2	751	40.7	3765	3549	5.01	4.73
WT_L1_C3	821	40.5	5526	5251	6.73	6.40
WT_L2_C3	786	40.7	7022	6732	8.93	8.56
WT_L3_C3	758	40.5	3861	3486	5.09	4.60
WT_L4_C3	767	40.6	6063	5835	7.90	7.61
WT_L1_C4	708	40.6	5017	4726	7.09	6.68
WT_L2_C4	750	40.6	7008	6693	9.34	8.92
WT_L3_C4	736	40.4	4301	3861	5.84	5.25
WT_L4_C4	737	40.7	7288	7090	9.89	9.62
EP_L1_CTRL	689	40.7	779	513	1.13	0.74
EP_L2_CTRL	767	40.6	905	590	1.18	0.77
WT_L1_CTRL	735	40.7	862	597	1.17	0.81
WT_L2_CTRL	740	40.4	900	617	1.22	0.83
TEM-1(Ancestor)	767	40.8	49	27	0.06	0.04
TEM-1(Ancestor)	800	40.6	50	31	0.06	0.04

3.5.3 SNPs found at frequencies above 10%

Table 3.3: Synonymous SNPs found at frequencies above than 10%. Rows are ordered according to the position (in Ambler numbering). SNPs found at frequency above 90% in at least one population are shown in bold.

				Frequency in cycle			
Position	SNP	Strain	Population	1	2	3	4
15	TTT15TTC	WT	1	47.6	0.0	0.0	0.0
21	**CTT21CTG**	EP	3	8.4	4.1	25.0	21.1
		WT	4	1.6	11.2	90.9	95.0
24	TTT24TTC	WT	1	46.1	0.0	0.1	0.0
76	**CTA76CTG**	WT	4	0.0	0.0	85.5	97.7
83	CGT83CGC	WT	2	0.0	22.6	4.2	0.0
84	GTT84GTC	EP	4	0.0	4.6	67.5	71.7
		WT	3	0.0	0.0	1.6	17.4
91	**CTC91CTT**	WT	1	0.2	99.4	97.8	97.7
97	TAT97TAC	EP	3	3.8	35.9	16.6	23.2
98	TCT98TCC	EP	3	6.5	3.8	24.1	18.7
107	CCA107CCG	WT	4	0.0	0.0	0.0	88.6
115	**GAT115GAC**	WT	4	0.0	4.5	86.3	98.8
120	AGA120AGG	EP	3	0.0	14.8	6.2	13.1
122	TTA122TTG	EP	2	11.5	5.8	0.4	0.3
	TTA122CTA	WT	2	0.1	24.9	4.1	0.0
144	GGA144GGG	WT	2	0.0	6.0	84.0	89.1
157	GAT157GAC	EP	2	0.0	13.0	0.1	0.6
162	**CTT162CTC**	WT	1	42.5	97.5	96.8	96.6
170	AAT170AAC	WT	2	0.0	19.6	0.0	0.0
184	GCA184GCG	EP	3	0.0	12.6	10.4	6.2
199	CTT199CTC	EP	1	0.0	36.8	0.2	0.2
207	TTA207CTA	EP	1	0.0	10.5	0.3	0.5
219	CCA219CCG	WT	2	0.0	15.1	0.0	0.0
225	CTT225CTC	EP	3	4.1	2.1	18.0	11.7
235	TCT235TCC	EP	4	1.2	2.2	63.4	73.7
274	GAA274GAG	EP	4	1.7	2.2	64.4	70.8
279	ATC279ATT	WT	3	0.0	20.8	13.7	2.6

Table 3.4: Nonsynonymous SNPs found at frequencies greater than 10%. Positions are given in Ambler numbering [156]. SNPs found at frequency above 90% in at least one population are shown in bold. SNPs known to have stabilizing effects are highlighted in cyan.

				Frequency in cycle			
Position	SNP	Strain	Population	1	2	3	4
13	**I13T**	WT	3	95.7	98.7	97.8	98.8
15	F15L	WT	2	0.0	0.0	77.1	85.7
16	F16L	WT	4	0.0	74.3	1.0	0.0
21	L21P	WT	4	77.4	75.9	3.1	0.0
34	**K34R**	WT	2	0.0	0.0	90.6	97.7
38	D38N	WT	3	0.0	19.3	16.9	19.3
47	**I47V**	EP	4	17.9	59.8	80.3	87.8
		WT	4	0.0	0.0	92.2	99.1
56	**I56V**	WT	2	98.3	98.2	98.7	97.5
100	**N100D**	WT	4	84.0	98.5	99.1	97.8
104	**E104K**	EP	1	0.0	98.2	98.1	97.4
			2	0.0	97.8	98.5	97.4
			3	0.0	98.2	97.6	98.0
			4	0.0	99.1	96.9	98.9
		WT	1	0.0	98.5	97.8	98.3
			2	0.0	98.3	97.3	98.8
			3	95.6	98.6	98.8	97.6
			4	0.0	98.7	98.0	98.4
112	H112Y	EP	1	0.0	10.2	0.0	0.0
			2	0.0	0.0	48.9	32.5
120	**R120G**	EP	4	10.0	12.7	4.4	4.6
140	T140A	EP	1	16.8	37.5	14.7	13.1
			2	12.2	10.7	10.8	6.9
141	T141A	EP	3	3.2	48.0	18.2	28.9
146	K146E	EP	1	0.5	10.9	0.0	0.3
147	E147G	EP	1	0.0	32.0	0.2	0.0
		WT	3	0.0	18.0	14.9	2.7
153	**H153R**	EP	3	0.0	3.4	37.8	25.9
		WT	3	0.0	41.0	68.7	96.9
	H153D	WT	1	20.3	0.0	0.0	0.0
154	N154S	WT	2	0.0	17.1	0.3	0.0
173	I173T	EP	2	0.0	12.7	0.1	0.1
182	**M182T**	EP	1	0.0	9.9	98.9	98.2
			2	8.8	98.7	72.7	95.7
			3	0.0	0.1	20.5	68.0
			4	0.0	3.0	80.9	90.3
		WT	1	38.8	98.9	97.8	99.6
			2	0.0	0.0	90.8	99.5
			4	0.0	0.0	87.1	98.8
208	**I208M**	WT	2	98.6	98.2	98.1	97.5
224	A224V	WT	3	0.0	0.0	13.1	26.1
238	**G238S**	EP	1	98.5	98.3	97.7	99.0
			2	98.6	99.4	97.9	97.7
			3	99.1	98.5	97.6	98.5
			4	98.1	99.3	97.6	98.5
		WT	1	98.6	99.3	98.4	98.9
			2	99.3	99.2	98.3	98.3
			3	98.6	98.5	98.0	98.1
			4	98.8	98.7	99.3	98.8
265	**T265M**	WT	1	20.9	99.6	99.0	98.9
268	S268G	EP	3	2.8	65.5	61.6	50.8
273	D273G	WT	4	0.0	0.0	0.0	82.9
288	**K288E**	WT	2	98.5	97.9	97.8	96.7

3.5.4 Primer sequences

Table 3.5: Primers and barcodes used for mutagenesis and sequencing

Primer	Sequence	Barcode
BC01	GGTAGGAGCAATGTAAAACGACGGCCAGT	AGCAAT
BC02	GGTAGGCCTGTTGTAAAACGACGGCCAGT	CCTGTT
BC03	GGTAGGGGGTTTGTAAAACGACGGCCAGT	GGGTTT
BC04	GGTAGGGAAGGCGTAAAACGACGGCCAGT	GAAGGC
BC09	GGTAGGTTAGGTGTAAAACGACGGCCAGT	TTAGGT
BC10	GGTAGGGTGCATGTAAAACGACGGCCAGT	GTGCAT
BC11	GGTAGGAACTTTGTAAAACGACGGCCAGT	AACTTT
BC12	GGTAGGGGATCGGTAAAACGACGGCCAGT	GGATCG
BC13	GGTAGGATAAGGGTAAAACGACGGCCAGT	ATAAGG
BC14	GGTAGGATTGGTGTAAAACGACGGCCAGT	ATTGGT
BC15	GGTAGGAGTGAGGTAAAACGACGGCCAGT	AGTGAG
BC16	GGTAGGCCCACCGTAAAACGACGGCCAGT	CCCACC
BC21	GGTAGGAACCTGGTAAAACGACGGCCAGT	AACCTG
BC22	GGTAGGCTTTGCGTAAAACGACGGCCAGT	CTTTGC
BC23	GGTAGGTGGAGAGTAAAACGACGGCCAGT	TGGAGA
BC24	GGTAGGAATTGTGTAAAACGACGGCCAGT	AATTGT
BC25	GGTAGGTGACGAGTAAAACGACGGCCAGT	TGACGA
BC27	GGTAGGGTTCAGGTAAAACGACGGCCAGT	GTTCAG
BC28	GGTAGGCTTCAAGTAAAACGACGGCCAGT	CTTCAA
TEM1FS-F	GTAAAACGACGGCCAGTGAATAATATTGAAAAAGGAAGC	-
TEM1FS-R	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGTAAACTTGGTCTGACAGGAGC	-
ELP	GGTAGGCAAGCAGAAGACGGCAT	-
TEM-F6	GCTTAAGAATAATATTGAAAAAGG	-
TEM-R6	GAATTGTAAACTTGGTCTGACA	-

Chapter 4

Characterizing mistranslation with mass-spectrometry proteomics

Abstract

Protein synthesis is a fundamental cellular process, yet it can be surprisingly error-prone. Errors in protein synthesis, called phenotypic mutations, can occur during transcription, tRNA charging, or translation. It is important to quantify and characterize phenotypic mutations because they have physiological and evolutionary consequences. Unlike genomic mutation rates, phenotypic mutation rates are difficult to measure directly. Our current estimates suggest that average mistranslation rates lie between 10^{-3} and 10^{-4} per codon. Current methods to measure these rates depend on indirect biochemical assays to detect mistranslation. These are laborious and can detect only a small subset of all possible mistranslation rates. The few available mass spectrometric (MS) analyses of mistranslation were focused on heterologously expressed proteins. Their observations might not reflect true physiological mistranslation rates. Here I use preexisting MS proteomics datasets to quantify proteome-wide mistranslation in two pathogenic bacteria. I use error-tolerant peptide-spectrum matching to identify mistranslated peptides in the MS data. I find that mistranslation occurs at high frequencies and introduces radical amino-acid changes into mistranslated proteins. I show that many essential proteins are subject to mistranslation. Furthermore, I find mistranslation affecting some proteins in identical positions across different conditions, and even in both of my two species. Most of these commonly mistranslated proteins have moonlighting functions in pathogenesis and virulence. These findings rise the intriguing possibility that mistranslation might be important in regulating bacterial pathogenesis.

4.1 Introduction

4.1.1 Biochemical noise and mistranslation

Biological systems are fraught with noise on all levels of organization due to the limited accuracy of biochemical processes [1]. Such noise results in phenotypic heterogeneity within populations, even when individuals in those populations are genetically identical [216, 217]. Biochemical noise can have both deleterious and beneficial consequences. Therefore, characterizing sources and the magnitude of noise is important.

Here, I focus on one instance of molecular noise, namely *mistranslation*. Mistranslation, or erroneous protein synthesis, occurs when ribosomes incorrectly decode mRNA and incorporate incorrect amino-acids into a growing peptide. Because of mistranslation, protein pools will contain subpopulations of proteins that differ in their amino acid sequence from the genes encoding the respective protein (figure 1.1). Protein pools affected by mistranslation are sometimes called *statistical proteins* [23]. Differences in sequence (i.e. mutations) between a

wild-type protein and mistranslated proteins are called *phenotypic mutations*. Since statistical proteins differ in sequence from wild-type proteins, they may differ in structure and function as well. On the one hand, phenotypic mutations can destabilize proteins, which causes protein misfolding and cytotoxic aggregation [22]. One other hand, statistical protein pools might contain functional diversity, facilitating adaptation to new environments [37].

4.1.2 Mistranslation rates

Bacterial genomic mutation rates have been the subject of many studies, and typically lie between 10^{-7} and 10^{-11} per base pair per generation [24]. While next-generation sequencing can help measure mutation rates, translational error rates are much harder to determine. In addition to difficulties in accurately determining sequences of rare protein variants in a cellular proteome, the space of possible mistranslation events is much larger than the space of possible mutations. Excluding the three stop codons, each of the 61 sense codons from the genetic table can be mistranslated into 19 different amino-acids. That means that there are 1159 (61x19) possible amino-acid misincorporations.

A variety of approaches have yielded estimates of average mistranslation rates between 10^{-5} and 10^{-3} per codon in *E. coli* and S. cerevisiae [16, 22, 26–28, 74, 218, 219], but these rates have been determined for only 5% of all possible amino acid misincorporations [22]. Importantly, estimated mistranslation rates vary by more than an order of magnitude across different codons [16, 28]. Surprisingly, average per-codon mistranslation rates in *Bacillus subtilis* and *Mycobacterium smegmatis* have been estimated to be as high as 10^{-2} in recent studies [37, 75]. With rates so high, error-free proteins would be virtually absent from a cell's protein pool.

4.1.3 Methods for measuring mistranslation rates

4.1.3.1 Indirect biochemical methods

Most studies used indirect biochemical methods to measure mistranslation rates [16, 22, 26–28, 74, 218]. These methods rely on using the activity of reporter genes as a proxy for determining mistranslation rates. One approach uses a dual-luciferase assay where the AAA lysine codon at position 529 in the firefly luciferase gene is mutated to a non-lysine codon. Because lysine at position 529 is crucial for the activity of luciferase, any luminescence produced from these luciferase mutants must come from misincorporation of lysine at position 529 during protein synthesis [16]. This method allows accurate measurements of mistranslation rates into lysine from all non-lysine codons. Recently, this approach was extended to other reporter genes, and measuring mistranslation rates to glutamate, aspartate, and tyrosine also became possible [28]. However, these indirect methods still have serious limitations. First, these methods only allow observations of a small subset of all possible mistranslation events. Second, finding a reporter gene for each of the 20 amino acids, and creating a mutant library that covers all codons is a laborious process. Third, one of the major factors determining the mistranslation rate is a codon's genetic context (e.g. neighboring codons) [99]. Measuring the effect of neighboring codons with the indirect approach would lead to a combinatorial explosion in the number of reporter gene constructs needed. Fourth, reporter genes used so far are heterologous to the host, and they are often overexpressed from a plasmid. It is therefore possible that mistranslation rates measured with these methods do not reflect the true physiological mistranslation rates.

4.1.3.2 Mass spectrometry proteomics

Mistranslation can be studied by directly detecting and sequencing mistranslated proteins through mass spectrometry (MS) proteomics. Generally, in "shotgun" MS proteomics the aim is to identify all proteins present in a sample using spectra that depend on a protein's mass [142]. Prior to the MS analysis, proteins in a sample are digested with specific proteases such

as trypsin to create a mixture of peptides to be analyzed. Experimentally determining peptide sequences using tandem mass-spectrometry (MS/MS) relies on isolating and recording the mass of a so-called *parent* or peptide ion in the first MS stage, and then fragmenting it and recording masses of resulting *fragment ions* in the second MS stage. When fragmentation occurs along the peptide backbone, fragment ion mass spectrum contains peptide "fingerprints" that can be used to infer the peptide's sequence [142].

In one approach, peptide sequences are assigned through a correlation of spectra acquired by MS/MS to spectra predicted from peptide sequences of the species under investigation [220]. These correlations, called *peptide-spectrum matches* (PSMs), can be assigned and scored using database searching algorithms such as X!Tandem [221], Sequest [222], and Mascot [223].

Conventional methods limit the search space by searching MS/MS spectra only against peptides whose masses match those spectra within a very narrow window. So-called *error-tolerant searches* test for known protein modifications by dynamically widening this window for each unexplained spectrum [224]. These error-tolerant searches also enable detection of changes in a peptide's sequence, i.e. through mistranslation, by allowing for mass-shifts that correspond to single amino acid substitutions in parent and fragment ions.

To limit the size of the search space, such error-tolerant searches are conducted in two stages. In the first stage, only a set of high scoring PSMs are generated, to identify peptides and proteins present in the sample with high confidence. From these PSMs, a new reduced database of theoretical peptide spectra is formed "on the fly". This second, reduced, database is used in the second, error-tolerant or *refinement*, stage to search for modified variants of peptides identified in the first search.

There are a few studies based on direct observations of mistranslated peptides from MS/MS proteomics experiments [98, 219], estimating mistranslation rates to $10^{-5} - 10^{-3}$ per codon. However, just like indirect methods, these studies measured mistranslation rates on overexpressed heterologous proteins.

4.1.4 Research aim

I wanted to fully characterize proteome-wide mistranslation across all codons in the genetic code. To this end, I used publicly available shotgun proteomics MS/MS data from two pathogenic bacterial species, *Escherichia coli* and *Shigella dysenterie* [225–228]. These data have been obtained from *in vitro* (laboratory cultivated) and *in vivo* (from infected animals) conditions, with three to four biological replicates per condition. I used error-tolerant searches with X!Tandem to identify mutated (mistranslated) peptides from these MS/MS spectra. In both species, I was able to identify mistranslation events for more than half of the genetic code.

Using a semi-quantitative method, I found that on average, mistranslation rates fall into ranges between 10^{-4} and 10^{-3}, as previously reported [16, 22, 28, 74]. However, I also found frequent ($\approx 10^{-2}$ per codon) mistranslation events that lead to radically different residues in proteins. This implies that phenotypic mutations have the potential to create a statistical pool of proteins with a significantly different biochemical profile than previously thought [16, 28].

I found many peptides that are mistranslated in identical positions across replicates and species. This opens the intriguing possibility that the ability to produce specific mistranslation products can be maintained by selection. Interestingly, commonly mistranslated proteins belong to a class of moonlighting proteins that carry out essential metabolic functions intracellularly, but when excreted are crucial for cell-to-cell signalling and virulence during infections. I discuss how these changes might affect the cellular physiology and virulence of pathogenic *E. coli* and *S. dysenterie*.

4.2 Materials and methods

4.2.1 Mass spectromety proteomics data

I used publicly available shotgun MS/MS proteomics data of two bacteria, enterohemorrhagic *E. coli* 86-24 [226, 228] and *Shigella dysenteriae* Sd197 [225, 227]. The experimental design for acquisition of the shotgun proteomics mass-spectrometric data for *E. coli* [226, 228] and *S. dysenteriae* [225, 227] was similar. Briefly, samples from both species were isolated from two conditions. Samples from *in vivo* conditions were isolated from guts of gnotobiotic piglets four days after infection, in three independent replicates (piglets) per species. Samples for *in vitro* condition were isolated from stationary phase cultures, in three replicates for *S. dysenteriae*, and four for *E. coli*.

I downloaded mass spectrometric data in the PrideXML format from the PRIDE PRoteomics IDEntifications (PRIDE) database [229]. Project IDs were PRD000502 (*in vitro*) and PRD000418 (*in vivo*) for *E. coli*, and PRD000500 (*in vitro*) and PRD000506 (*in vivo*) for *S. dysenteriae* Sd197. I used jmzReader [230] to convert spectral data from the PrideXML format to MASCOT generic format (.mgf), which could then be analyzed by X!Tandem (section 4.2.3).

4.2.2 Construction of the target-decoy database

I downloaded the protein and genomic sequence data for *E. coli* and *S. dysenteriae* from GenBank (date of download Oct 12 2015). At the time of the analysis, a high quality sequence and annotated genome did not exist for *E. coli* 86-24. Instead I used the genomic sequences of a closely related strain *E. coli* EDL933, the same as used in the publications of the *E. coli* data [226, 228].

Typically, searching a large number of spectra, e.g. from a whole-proteome shotgun MS/MS analysis, against a sequence database of the entire proteome results in a large number of PSMs, some of which might be false. A common problem is to validate PSMs, and to estimate the rate of false peptide and protein identifications [231]. Thus, I wanted to construct a database that would enable me to estimate the rate of false peptide identifications. This is commonly done by concatenating a database of protein sequences coming from the proteome of interest (*targets*) and a database of proteins whose amino acid sequences are randomized (*decoys*) [231]. When decoy sequences are constructed properly, the number of false identifications will be evenly distributed among target and decoy peptides. If the sizes of target and decoy databases are the same, the false discovery rate (FDR) is simply the ratio between decoy and target PSMs, because all decoy identifications are false. Thus, the correct estimate of FDR critically depends on knowing the ratio of target to decoy peptides in the database.

Because identifying mistranslated peptides involves two successive searches, estimating false discovery rates is more complicated. In the first search, only a set of high quality PSMs are generated. From these PSMs, a new reduced database of proteins is formed ”on the fly”. This second, reduced, database is used for refinement of the search, i.e. to search for variants of peptides that are modified (chemically, or through post-translational modification and mutation/mistranslation). Because target peptides are more likely to be identified than decoys in the first search, the second database might have a distorted target-to-decoy ratio, which makes the estimation of false discovery rates (FDR) impossible. To circumvent this issue, I constructed my search database as follows. I compiled two separate peptide sequence databases, one for *E. coli*, and one for *S. dysenteriae*. To each of the protein sequences in the database, I appended the reversed sequence of the same protein (figure 4.1). Thus, the first half of each protein sequence in the database produces target peptide sequences, while the second half produces decoy peptide sequences. This approach ensures that: a) the decoy database has the same amino acid composition as the target database, and b) that the same ratio of targets to decoys is present in both the first and the refinement stage of the PSM search. I classified a PSM as a target hit if the identified peptide mapped to the first part of a

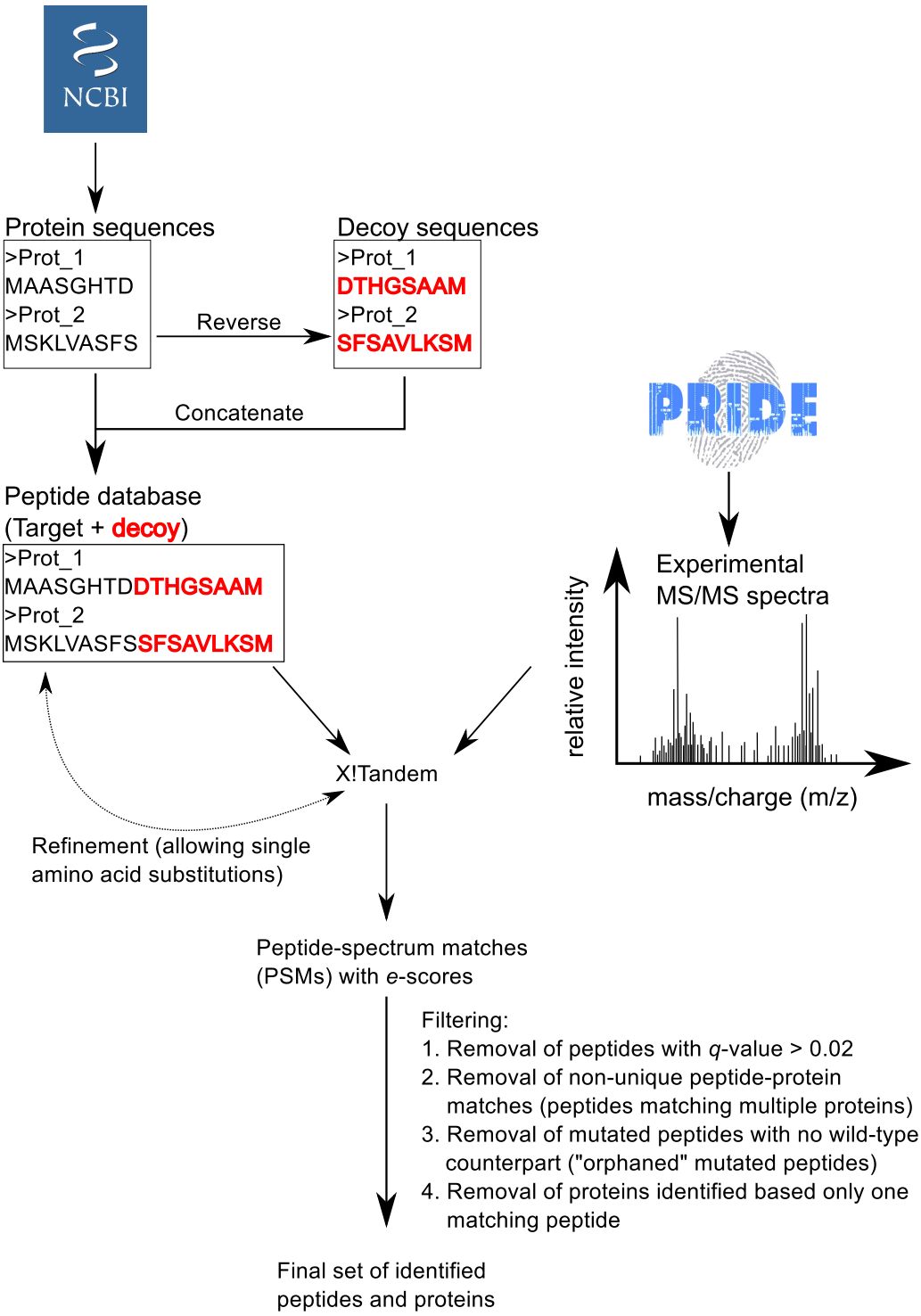

Figure 4.1: Method used to identify and filter mutated (mistranslated) peptides in MS/MS spectra. I used the X!Tandem search engine to identify peptides in publicly available MS proteomics data from *E. coli* and *S. dysenteriae*. I used the error-tolerant mode of X!Tandem (refinement by allowing single amino acid substitutions) to identify mistranslated peptides in concatenated target-decoy peptide databases. I quality-filtered identified peptides to generate a final set of wild-type and mistranslated peptides and proteins for my downstream analyses of mistranslation.

protein sequence in the database, and as a decoy hit otherwise.

To account contaminants commonly identified in MS proteomics experiments, I downloaded sequences of common contaminant peptides from the Global Proteome Machine ftp site (ftp.thegpm.org/fasta/cRAP) on Oct 12 2015. I appended the reversed sequence to contaminant peptides as described above. I then used the combined database of bacterial and contaminant peptides as the final database for my PSM search.

4.2.3 Analysis of MS/MS data

In my analysis I used X!Tandem [221], a software package that can simulate theoretical spectra from peptide databases and match them to experimental spectra. In short, I used X!Tandem to search the whole-proteome shotgun MS data from *E. coli* and *S. dysenteriae* against the target-decoy peptide database (section 4.2.2, figure 4.1). I set the mass error tolerance to ±0.4 Da for fragment ions, and to ±1.5 Da for parent ions. I used cysteine alkylation due to iodoacetamide (+57.022 Da) as a fixed modification in all searches, and allowed for oxidation of methionine (+15.995 Da) and carbamoylation of lysine (+43.01 Da) as potential modifications. I performed the search in an error-tolerant mode (i.e. with refinement), allowing for one missed cleavage, as well a single amino acid change relative to the expected peptide (to identify mistranslated peptides). In both main and refinement searches, I used an *e*-value of 0.001 as a threshold value for accepting peptide identifications.

4.2.4 False discovery rates

I calculated false discovery rates (FDR) and *q*-values as described in [232]. The FDR is calculated as follows:
The number of false positives (FP) is the number of decoy hits. I subtracted this number from the number of all hits above threshold *e*-value to get the number of true positives (TP). I then calculated the FDR from FP and TP as

$$FDR = \frac{FP}{TP + FP} \tag{4.1}$$

I used the following algorithm to calculate FDR and a *q*-value for each peptide-spectrum match [232]:

1. I sorted all PSMs according to their *e*-values
2. I traversed the ordered PSM list from the lowest to the highest *e*-value. For each PSM, I calculated the cumulative FDR according to the formula (4.1), taking into account all identified true (target) and false (decoy) positive peptides up to that point, and stored this value as FDR_{est}.
3. I traversed the ordered PSM list from the highest to the lowest *e*-value, storing the minimal FDR value observed up to that point (FDR_{min}). I then assigned a *q*-value, which is the minimum FDR at which the peptide identification could be made, to each PSM in the following way. If FDR_{est} was greater than FDR_{min} for a peptide, I assigned *q*-value for that peptide to be equal to FDR_{min}. If FDR_{est} was lower than FDR_{min} otherwise I assigned I assigned *q*-value to be equal to FDR_{est}, and set FDR_{min} to FDR_{est}.

4.2.5 Peptide-to-spectrum match quality filtering

I employed a set of criteria to improve the confidence of my peptide identifications (figure 4.1). First, I set the false discovery rate to 2%, and removed PSMs whose *q*-values were above 0.02. To increase the confidence of peptide-to-protein matches, I removed all peptides that could be matched to multiple proteins, and those proteins that were identified based on only one peptide.

Additionally, I wanted to distinguish between amino acid substitutions resulting from mistranslation and those coming from genetic mutations. To increase the confidence that

modified peptides are really mistranslated, I removed all "orphaned" mistranslated peptides, i.e. I considered peptides with single amino acid changes to be the result of mistranslation only if the wild-type peptide was identified as well.

4.2.6 Estimating mistranslation frequencies

I used a semi-quantitative method to estimate per-codon mistranslation frequencies. First, I mapped all residues of identified peptides to their corresponding codons in the genomic nucleotide sequences. For each of 61 sense codons in the genetic code, I recorded the number of times it was correctly and incorrectly translated. Specifically, I counted the codon as correctly translated if the residue in the peptide is the same as one would expect based on the standard genetic code, and as mistranslated otherwise. I kept record of each codon-to-amino-acid mistranslation separately. I assumed that a specific mistranslation event is biologically important if it can be detected independently in multiple independent samples (replicates) of the same species. Therefore, I considered a mistranslation event to be a true positive for a condition (*in vivo* or *in vitro*) only if it was observed in all independent samples from that condition. I calculated the mistranslation frequency (MF) at a codon for all 61 sense codons as:

$$MF_{COD} = \frac{M_{COD}}{N_{COD}} \tag{4.2}$$

where M_{COD} is the number of times a mistranslation event has been observed at a specific codon, and N_{COD} is the total number of times that codon was observed in the dataset, i.e. the number of times a residue in an identified peptide was mapped to that codon. I considered both aggregated and individual mistranslation frequencies. I calculated the aggregated mistranslation frequency, a measure of "error-proneness" of a codon, by grouping all mistranslation events at a single codon, regardless of the resulting amino-acid. In contrast, individual mistranslation frequencies quantify how often a particular codon-to-amino-acid mistranslation event occurs, and I estimated them by considering each mistranslation type separately. For example, consider a hypothetical dataset of 990 instances of error-free peptides that contain Phe encoded by TTC. In the same dataset, there are 7 instances of a mutated peptide containing Ser instead of Phe; and 3 instances of a mutated peptide with Leu instead of Phe. The aggregated MF_{TTC} is $(7+3)/(990+7+3) = 10^{-2}$, whereas individual mistranslation frequencies are $MF_{TTC \to Ser} = 7/(990+7+3) = 7 \times 10^{-3}$, and $MF_{TTC \to Leu} = 3/(990+7+3) = 3 \times 10^{-3}$.

I was interested in functions and abundances of proteins that were commonly mistranslated between *E. coli* and *S. dysenteriae*. To this end, I retrieved the *E. coli* protein abundance from the PaxDB database 4.0 [233], and I generated a list of moonlighting protein functions from MultitaskProtDB, the database of moonlighting proteins [234]

4.3 Results

I wanted to estimate mistranslation rates for all sense codons in the genetic code. To this end, I used publicly available whole-proteome MS shotgun proteomics data from enterohemorrhagic *E. coli* 86-24 [226, 228] and *Shigella dysenteriae* Sd197 [225, 227]. Briefly, I used error-tolerant searches with X!Tandem search engine [221] to identify wild-type and mistranslated peptides. I then filtered identified peptides based on several criteria (figure 4.1) to produce a final set of peptides and proteins for downstream analyses. I used counts of mistranslated and wild-type peptides to estimate of per-codon mistranslation rates.

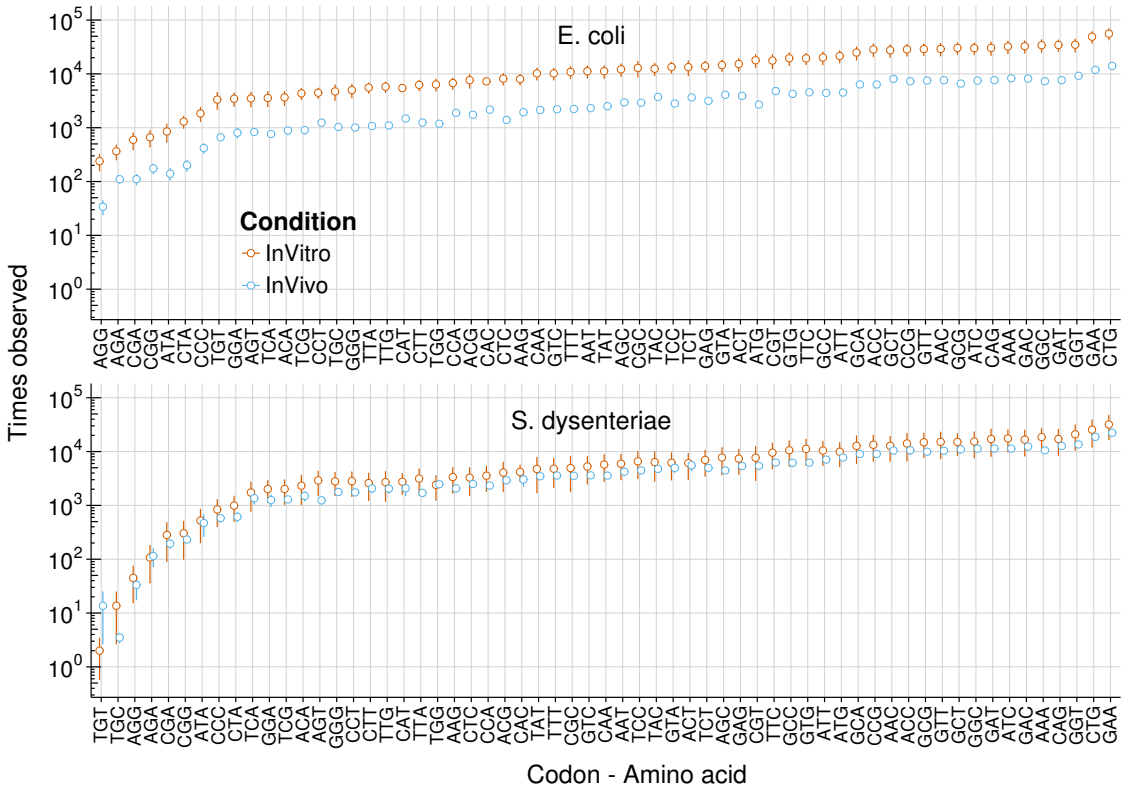

Figure 4.2: Codon coverage in MS/MS protemic datasets. I mapped peptides identified in the spectral search to their respective coding sequences in reference genomes of *E. coli* EDL933 and *S. dysenteriae* Sd197. Circles represent mean numbers of codon observations per species and condition (*in vivo* or *in vitro*). Error bars represent standard deviations of the observed codon count across three biological replicates per condition.

4.3.1 Coverage of the genetic code by MS/MS datasets varies by more than two orders of magnitude

My results show that shotgun MS/MS proteomics data can be used to detect and estimate mistranslation at a coverage that is typically unattainable through indirect biochemical methods. I mapped residues of all identified peptides to codons they were translated from, and recorded how many times each of the 61 sense codons is observed. When examining how many times a codon is observed in my data set, I found that the per-codon coverage varies by two orders of magnitude across the genetic code, with most codons being covered between 10^3-10^4 times (figure 4.2). This enables detection of mistranslation events which occur at a rate of 10^{-4} per codon or higher. The variation in coverage can be explained by the biased used of synonymous codons in genes coding for highly expressed genes.

4.3.2 Mistranslation frequencies vary by two orders of magnitude between codons

Next, I estimated mistranslation frequencies from the identified wild-type and mistranslated peptides. I use the high per-codon coverage, as well as three to four biological replicates (independent samples of the same species in the same condition) to increase the confidence of my estimate. Specifically, I only consider a codon to be mistranslated if it has been observed as mistranslated at least twice, in at least two distinct peptides, and more than two independent replicates. I calculated the frequency of mistranslation for a codon as the ratio of the number of times the codon was mistranslated to the total number of times residues in identified

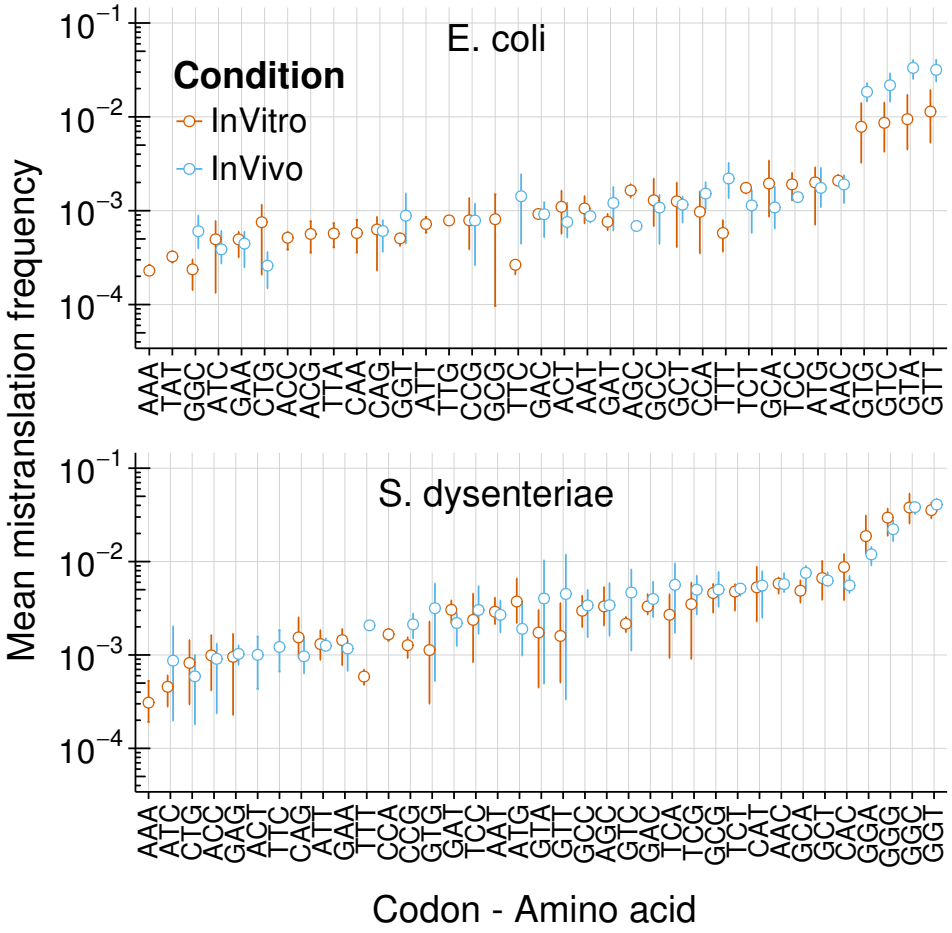

Figure 4.3: Average per-codon mistranslation frequencies. I grouped all the observed mistranslation events according to the codon they mapped to and calculated the mistranslation frequency at a codon as the ratio of the number of times a codon was mistranslated to the total number of times that codon was observed. I calculated mean mistranslation frequencies separately for *in vivo* and *in vitro* conditions for both species. Circles are means, and error bars show standard deviations across biological replicates from a condition.

peptides were mapped to that codon. The procedure described in section 4.2.6 yielded mistranslation frequency estimates at 35 codons for *in vitro* conditions for both species, and at 26 and 35 codons for *in vivo* conditions, in *E. coli* and *S. dysenteriae*, respectively. My estimated average per-codon rates of mistranslation vary between 10^{-2} and 10^{-4} per codon (figure 4.3), in concordance with previously published mistranslation rates [16, 22, 26–28, 37, 74, 75, 218, 219]. Table 4.1 lists the most robust (the lowest mistranslation frequency) and the most erroneous (the highest mistranslation frequency) synonymous codons for all amino acids encoded by multiple codons in *E. coli* and *S. dysenteriae*.

4.3.3 Mistranslation frequently leads to radical amino acid substitutions

I wanted to characterize mistranslation with respect to the changes in amino acid composition it creates. To this end, I considered each codon-to-amino acid mistranslation individually. I grouped the substituted amino-acids into one-mutant neighbors and non-one-mutant neighbors of the wild-type amino-acid, according to the standard genetic code. I considered substituted amino-acids to be one-mutant neighbors, if any of their codons could result from a single mismatch to any of the codons encoding the wild-type amino-acid, i.e. if mistranslation could occur by a single mismatch in codon-anticodon matching on the ribosome. The most

Table 4.1: The most accurate and the most error-prone synonymous codons in *E. coli* (EC) and *S. dysenteriae* (SD). I ranked synonymous codons based on their mistranslation frequencies, and selected the most accurate and the most error-prone codon for an amino acid. Methionine and tryptophan are omitted, since they are encoded by a single codon. Arginine and cysteine are omitted since I detected no mistranslation events at codons for either amino acid.

Amino Acid	Accurate codon		Error-prone codon	
	EC	SD	EC	SD
A	GCG	GCC	GCA	GCT
D	GAT	GAT	GAC	GAC
E	GAG	GAG	GAA	GAA
F	TTT	TTT	TTC	TTC
G	GGG	GGA	GGT	GGC
H	CAC	CAT	CAT	CAC
I	ATA	ATA	ATC	ATT
K	AAG	AAG	AAA	AAA
L	CTA	CTT	CTG	CTC
N	AAT	AAT	AAC	AAC
P	CCC	CCC	CCA	CCA
Q	CAA	CAA	CAG	CAG
S	TCA	AGT	TCT	TCT
T	ACA	ACA	ACT	ACC
V	GTG	GTG	GTT	GTT
Y	TAC	TAT	TAT	TAC

Table 4.2: Ten most frequent mistranslation events in *E. coli* data sets. I calculated the mistranslation frequency by aggregating all identified peptides from all samples obtained from *E. coli*.

Codon	Mistranslation	Mistranslation frequency $\times 10^{-3}$	Times observed mistranslated
GTT	V→H	15.79	1712
GTA	V→H	15.24	861
GTC	V→H	10.85	411
GTG	V→H	9.28	681
AAC	N→P	1.92	222
TCC	S→D	1.92	60
ATG	M→W	1.78	14
TCT	S→D	1.67	88
AGC	S→D	1.48	40
GCA	A→N	1.17	78

Table 4.3: Ten most frequent mistranslation events in *S. dysenteriae* data sets. I calculated the mistranslation frequency by aggregating all identified peptides from all samples obtained from *S. dysenteriae*.

Codon	Mistranslation	Mistranslation frequency $\times 10^{-3}$	Times observed mistranslated
GGC	G→P	39.90	3209
GGT	G→P	36.73	3634
GGG	G→P	26.78	352
GGA	G→P	18.10	164
CAC	H→Y	6.33	118
CAT	H→Y	4.79	64
GCT	A→P	4.45	335
ATG	M→Y	4.25	29
GTA	V→H	4.14	92
AAC	N→P	3.94	281

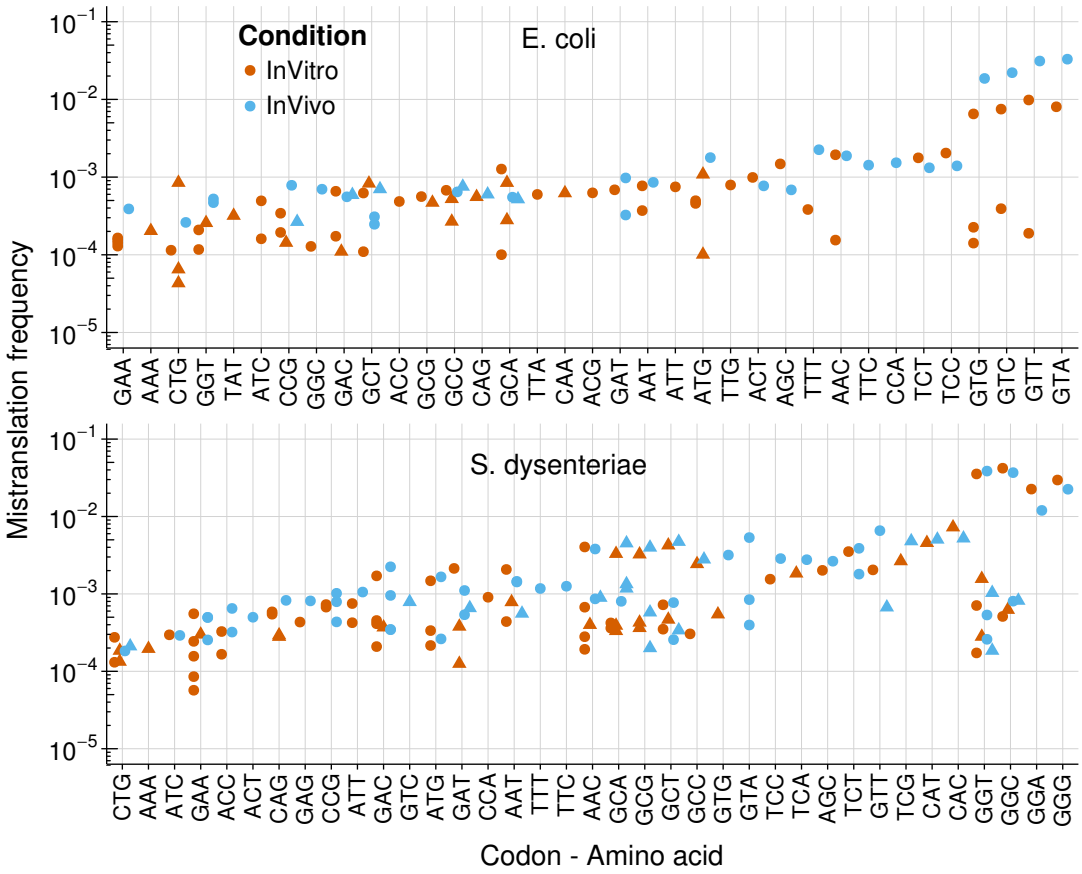

Figure 4.4: Frequencies of individual codon-to-amino acid mistranslations. I grouped all the observed mistranslation events according to the codon they mapped to, and according to the amino acid substitution they cause. If a substituted amino acid can be caused by a single mismatch between the codon and anticodon during translation, I would classify the event as a one-mutant neighbor mistranslation (triangle), otherwise I would classify it as a non-one-mutant neighbor mistranslation (circle)

frequently observed mistranslation evens are not of the one-mutant neighbor kind (figure 4.4, tables 4.2 and 4.3). For example, most commonly observed mistranslations in *E. coli* are valine-to-histidine changes (table 4.2), while the most common mistranslation for *S. dysenteriae* is glycine-to-proline (table 4.3).

4.3.4 Mistranslation affects many proteins

I wanted to find out if the mistranslation events I observe come from erroneous translation at few ”mistranslational hot-spots”, or if they are distributed over many different loci. To this end, I examined how many distinct peptides are affected by each of the mistranslation events shown in figure 4.4, and found high frequency mistranslation in tens of distinct peptides (figure 4.5). The highest frequency mistranslation events (V→H and G→P) (tables 4.2 and 4.3) are also those that occur at the highest number of distinct peptides.

4.3.5 Mistranslation can be conserved across conditions and species

Mistranslation is by definition stochastic. However, because it is determined by factors such as codon identity and genetic context, some mistranslation products might be reproducibly synthesized by cells in different conditions. To find out to what extent mistranslation events happen at the same loci, and give rise to the same peptides, I examined peptides that are identically mistranslated across different independent replicates and culture conditions (*in vivo*

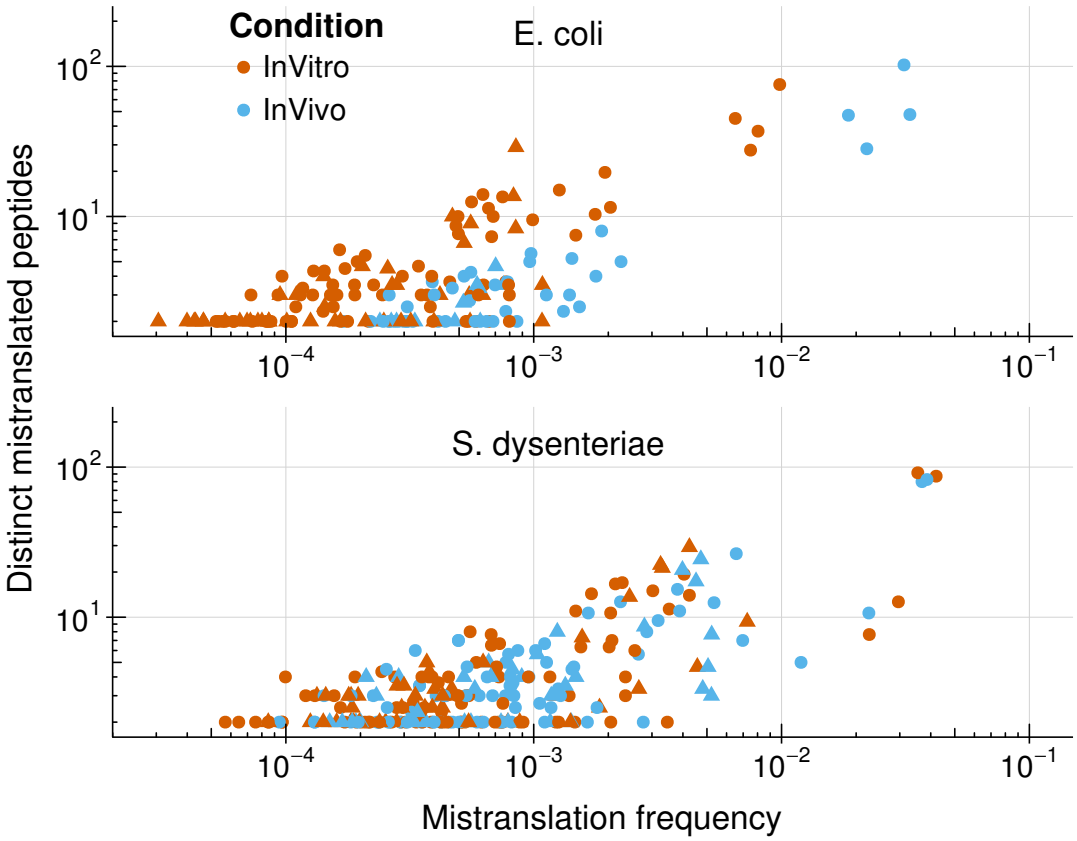

Figure 4.5: Mistranslation affects many different peptides. Each data point relates a frequency of a specific codon-to-amino acid mistranslation event to the number of distinct peptides it affects. Triangles depict mistranslation to amino acids that are one-mutant neighbors of the wild-type amino acid in the genetic code. Circles show mistranslation events that are not one-mutant neighbors.

vs *in vitro*) in the same species. I found 32 instances of peptides identically mistranslated in all *E. coli* samples (top 10 most frequently mistranslated shown in table 4.5), and 50 in all *S. dysenterie* samples (top 10 most frequently mistranslated shown in table 4.4).

E. coli and *S. dysenterie* are genetically very close species [235], with many genes showing high sequence conservation on the nucleotide (codon) level. Therefore, I wanted to find out if any of peptides are identically mistranslated in these two species across multiple replicates. For this analysis, I considered replicates from *in vivo* and *in vitro* conditions together. I found 8 instances of peptides (table 4.6) identically mistranslated in at least three samples of each of *E. coli* and *S. dysenterie*. I found only one protein to be mistranslated identically in at least five replicates per species. Specifically, the DNA-binding protein HLP-II (H-NS) was mistranslated from glutamic acid (GAA) to tyrosine (TAT/C) in position 63.

4.4 Discussion

The flow of genetic information from genes to functional proteins is error-prone. One of the largest sources of errors during this process is translation. While high rates of mistranslation cause fitness defects and disease [22], recent studies hint at possibility that in certain species or environmental conditions, mistranslation could have an adaptive role [41, 132, 236]. Accurate and comprehensive measurements and characterization of translational errors are necessary to understand the evolutionary consequences of mistranslation.

Traditionally, indirect biochemical methods were used to measure mistranslation rates [16, 22, 26–28, 74, 218]. However, these methods are limited in their scope. Here I demonstrate

Table 4.4: Top 10 most frequently mistranslated proteins and their codons found identically mistranslated in all *S. dysenteriae* samples.

Protein ID	Codon	Mutation	Product
558576524	GGC	G62P	GroES
558576525	GGT	G382P	GroEL
558572105	GGT	G172P	Cystine-binding protein
558575627	GGG	G857P	RNA polymerase β chain
558572671	GGT	G301P	Glyceraldehyde 3-phosphate dehydrogenase
558575289	GGT	G57P	Ribosomal protein L6P
558574438	GGT	G57P	Ribosomal protein S16P
558573849	GGC	G412P	Ribosomal protein S1P
558574879	GGT	G25P	Transketolase
558577052	GGC	G147P	Type III secretion system effector protein IpaC

Table 4.5: Top 10 most frequently mistranslated proteins and their codons found identically mistranslated in all *E. coli* samples.

Protein ID	Codon	Mutation	Product
12519118	GTC	V98H	Aspartate ammonia-lyase
12519125	GTT	V464H	GroEL
12519125	GTA	V94H	GroEL
12518575	GTT	V75H	High-affinity phosphate-specific transport system
12514142	GTA	V199H	Outer membrane protein 3a
12517506	GTG	V306H	Periplasmic L-asparaginase II
12517020	GTT	V60H	Putative yhbH sigma 54 modulator
12512824	GTG	V206H	Pyruvate dehydrogenase
12517444	GTT	V50H	Ribosephosphate isomerase
12512683	GTG	V15H	Transaldolase B

Table 4.6: Proteins and codons identically mistranslated in both *E. coli* and *S. dysenteriae*. The abundance of a protein is expressed as the abundance rank among *E. coli* proteins in the PaxDB database [233].

Codon	Mutation	Protein	Abundance rank
GAC	D60F	Ribosomal protein L6	74/4096
GAT	D460P	DnaK (Hsp70)	17/4096
AAC	N103P	DNA-binding protein HLP-II (H-NS)	9/4096
GAA	E63Y	DNA-binding protein HLP-II (H-NS)	9/4096
GAC	D68P	DNA-binding protein HLP-II (H-NS)	9/4096
AAC	N72P	Enolase	25/4096
GAT	D334P	GroEL	39/4096
GTC	V82M	Isocitrate dehydrogenase	63/4096

that MS-based proteomics can be used to estimate mistranslation rates as low as 10^{-4} per codon. Furthermore, I show that mistranslation rates can be higher than 10^{-2} per codon, and that mistranslation often creates radical changes in the proteome. Some of these mistranslation events occur at identical positions in datasets from multiple independent replicate samples, across two experimental conditions (*in vivo* and *in vitro*), and in two different species, implying evolutionary conservation and suggesting an adaptive role for mistranslation.

4.4.1 Feasibility of using MS/MS-based shotgun proteomics to quantify mistranslation

The only previous study to use MS/MS proteomics for measuring mistranslation in proteins endogenous to the host was focused on serum albumin in humans [219]. However, because mistranslation can vary greatly depending on the level of protein expression and the sequence context [99], mistranslation needs to be studied for a large number of genes. I demonstrate that it is possible to study mistranslation across majority of sense codons from the genetic code

using whole-proteome MS/MS-based shotgun proteomics. I was able to identify mistranslation events for up to 35 sense codons in *E. coli* and *S. dysenteriae*. However, my approach is limited to identifying mistranslation rates higher than 10^{-4} per codon (figure 4.2). Furthermore, it is biased towards highly expressed genes, and these genes infrequently contain rare codons.

4.4.2 Mistranslation rates

I find that mistranslation occurs at a rate that is on the order of $10^{-4} - 10^{-3}$ per codon for the majority of codons in the genetic code (figure 4.3). My estimates are in agreement with previously published data on mistranslation [16, 22, 26–28, 74, 218, 219]. I also consistently observe high frequency ($\approx 10^{-2}$ per codon) mistranslation at some codons. Mistranslation rates as high or higher than 10^{-2} are not unheard of, and seem to be tolerated by cells [75, 98, 131, 237]. I observe no mistranslation rates lower than 10^{-4}, although previous studies report the lowest measured mistranslation rates to be about 2×10^{-6} [28]. This discrepancy reflects my inability to detect rare mistranslation events, and to detect and confidently estimate mistranslation at rare codons.

It is unlikely that the high frequency of mistranslation I observe is due to chemical artifacts. First, according to the unimod, a database of known protein modifications [238], no known post-translational modifications or MS sample artifacts cause mass-shifts identified by X!Tandem as mistranslation. Second, all samples were prepared and analyzed by mass spectrometry using the same protocol [225–228], so one would expect similar chemical artifacts to occur in all cases, but samples from *E. coli* have high levels of V→H, while samples from *S. dysenteriae* have high levels of G→P mistranslation. Finally, while the mass-shifts I observed could in principle caused by mutations occurring in exponential and stationary phase, mutation is an unlikely explanation for high frequency mistranslation in my data for four reasons. First, I ruled out preexisting mutations in the bacterial stock by only considering mutated peptides if they have a wild-type counterpart in the sample. Second, mistranslation events, and the exact positions where they occur, are shared across biological replicates from the same species (tables 4.4 and 4.5), and sometimes even across samples from different species (table 4.6). Third, high frequency mistranslation events are found among many different peptides (loci) (figure 4.5). Fourth, many of the mistranslated proteins are essential, and radical amino-acid substitutions are more likely to be detrimental if they are genetic than if they only occur through mistranslation and thus affect only some of the protein pool.

For the majority of amino acids, the most accurate and the most error-prone synonymous codons are often identical in *E. coli* and *S. dysenteriae* (table 4.1). However, there are some codons that greatly differ in their propensity to be mistranslated among the two species (figure 4.3 and table 4.1). Although these two bacterial species show high sequence conservation [235], there are other factors that might contribute to the differences in their per-codon mistranslation rates. For example, different copy numbers of tRNA genes are known to affect tRNA concentration in bacteria [239], and this in turn affects mistranslation rates at cognate and near-cognate codons [16]. In addition, differences in tRNA modifications can influence accuracy of decoding [173]. Differences in most accurate codons among *E. coli* and *S. dysenteriae* suggest that universally accurate/error-prone codons might not exist. Instead, per-codon mistranslation rates are likely species-specific, and need to be measured across different organisms.

4.4.3 Biological consequences of mistranslation

I found two surprising results in this study. First, among all mistranslation events, there are many that lead to radically different amino-acids (figure 4.4). This is unexpected because single mismatches in codon-anticodon pairing are thought to be the main drivers of mistranslation [16, 23]. Furthermore, among three possible positions, mismatches are expected to happen at high rates only in the third position of the codon-anticodon pair [23] because this results in

substitutions for amino acids with similar physicochemical properties, and limits the damaging effects of mistranslation.

In contrast to this expectation, I find that most frequent changes can not be explained by a single mismatch in codon-anticodon pairing. These changes are V→H changes in *E. coli*, and G→P in *S. dysenteriae* (figure 4.4, tables 4.2 and 4.3). Both of these types of changes occur at a frequency higher than 10^{-2}.

Valine to histidine changes involve substituting an aliphatic residue with a positively charged one. This can cause disruption of the protein structure. Changing glycine to proline does not affect the polarity (i.e. both amino acids are non-polar), but can nevertheless have catastrophic consequences on the protein structure. The reason is that glycine is a small amino acid with no side chain, and offers flexibility to the protein backbone, while proline offers very little flexibility and can disrupt α-helices [240, 241]

In contrast to being potentially damaging, these types of substitutions can also help create proteins with diverse biophysical and biochemical properties. For example, ambiguous mistranslation of CUG as either serine or leucine in *Candida albicans* gives rise to surface proteins with heterogeneous physical properties. This can affect adhesion during tissue invasion, or create surface variations that enable host immune system evasion [35].

A second important observation is that peptides identically mistranslated in *E. coli* and *S. dysenteriae* (table 4.6) map to proteins that often carry out essential functions in the cell, and that they are all in the top 5% of the most abundant proteins in the *E. coli* proteome [233]. That highly expressed essential proteins are reproducibly found mistranslated means that cells are in principle more robust to mistranslated proteins than previously thought. Moreover, mistranslated variants might be conserved because phenotypic mutations impart new functions to them. Some of these mistranslated proteins can be classified as "moonlighting proteins". Moonlighting proteins are a class of proteins that perform multiple functions without having distinct protein domains associated with these functions [242]. Potential examples where even a single amino-acid change can alter a protein's structure and function include proteins called "neomorphic moonlighting proteins" [243]. Orthologs of five out of six commonly mistranslated proteins in my data set (table 4.6) are involved in virulence and pathogenicity, or have moonlighting functions involved in virulence and infection in various pathogenic bacteria. Moonlighting functions of these proteins might be facilitated by mistranslation.

First, DnaK (Hsp70) is a chaperone whose numerous moonlighting functions, such as plasminogen binding in *Mycobacterium tuberculosis* [244] might be promoted through mistranslation. Second, the metabolic enzyme enolase is found on the cell surface of pathogens such as *Borrelia burgdorferi* [245], as well as pathogenic streptococci and staphylococci, as an adhesion factor that binds plasmin [246], plasminogen [246, 247], and laminin [248]. Adhesive properties of enolase might be enhanced by mistranslation, similar to surface proteins in *C. albicans* [35]. Third, GroEL is a molecular chaperone with many reported moonlighting functions [249]. Most importantly, it is involved in pathogen-host interactions by modulating cell adherence [250], cell invasion [251], and proteolytic activity in *Mycobacterium leprae* [252]. Intriguingly, there is evidence that *E. coli* GroEL can acquire different biological functions through a small number of amino acid changes [253], making it a good candidate gene for neofunctionalization through mistranslation . Fourth, isocitrate dehydrogenase can be used as a virulence factor in *Helicobacter pylori* [254], and in humans it can gain novel biochemical activity and physiological functions through single amino-acid changes [243]. Finally, DNA-binding protein H-NS, although it has no reported moonlighting functions, might regulate cellular physiology through mistranslation. H-NS is a DNA binding protein similar to histones that regulates pathogenicity [255], secretion of protein toxins [256], and expression of horizontally acquired genes, which are often involved in virulence [257, 258]. Mistranslating H-NS could create variants that exhibit different regulatory properties, because mutational analysis of H-NS reveals many substitutions that affect its DNA binding or the ability to transcriptionally repress other genes [259]

Of six commonly mistranslated proteins in *E. coli* and *S. dysenteriae*, ribosomal protein L6 is the only not implicated in virulence. However, mistranslated variants of L6 might affect cell's physiology by modifying the rate of mistranslation. Mutations of this protein can increase ribosomal fidelity, as well as resistance to streptomycin and gentamycin [260]. This observation raises an intriguing possibility, namely that mistranslation might produce erroneous protein variant that can reduce the rate of mistranslation, thus creating a feedback loop for regulating the fidelity of translation. It is also possible, however, that L6 might be involved in unknown moonlighting activities, since most other ribosomal proteins have moonlighting functions in replication [261], transcription [262, 263], RNA processing [264], DNA repair [265], and regulation of translation [263]. Any of these activities might depend on or be facilitated by mistranslation.

4.4.4 Limitations and future studies

My aim was to comprehensively estimate mistranslation rates for all sense codons in the genetic code using MS proteomics. However, I was able to do so for only about 35 out of 61 codons. My approach was limited by four factors.

First, in whole-proteome shotgun MS proteomics most detected peptides come from highly abundant peptides and proteins. Because highly expressed proteins are typically translated from frequent codons, conventional MS proteomics can not be used to measure mistranslation rates on rare codons (figure 4.2).

Second, the most highly abundant codons in these data are observed fewer than 10^5 times (figure 4.2). Given that I consider mass shifts to be true mistranslation events only if they are observed multiple times, the lower limit of detection for mistranslation rates from whole proteome shotgun MS data is around 10^{-4}. In contrast, indirect biochemical methods have a lower limit of detection around 10^{-6} per codon [28].

Third, previous studies show that mistranslation rates are affected by the genetic context, mainly by neighboring codons [98, 99]. To completely characterize mistranslation, it is necessary to measure mistranslation rates of the same codon in multiple contexts. However, the limited number of observations in my datasets, and the rareness of mistranslation events makes this task impossible with conventional shotgun proteomics.

The fourth limitation is not specific to this study, but rather affects all MS proteomics approaches to detect mistranslation. Mass-shifts can be validated to be true amino-acid substitutions in peptide-spectrum matches, by using synthetic peptides with engineered substitutions, and comparing their mass spectra with the spectra of putative mistranslated peptides [266]. However, this does not prove that the substitution results from mistranslation. Rather, it can result from genomic mutations, mistranscription, or mischarging of tRNAs. To validate that mass-shifts really result from mistranslation, one would need to carefully design an experiment that alters mistranslation rates, but not rates of other errors. This could be done by introducing ribosomal mutations into experimental strains, or by using antibiotics that increase only mistranslation rates, such as streptomycin.

With further advances in mass-spectrometry proteomics, a comprehensive characterization of mistranslation rates, and all factors that affect them, might become possible in the near future.

Chapter 5

Conclusion

Protein synthesis, despite being critical for cellular function, is remarkably error-prone. When ribosomes incorrectly decode mRNA, they introduce phenotypic mutations into the nascent peptide. Phenotypic mutations affect the structure and function of proteins. Because mistranslation has real physiological and evolutionary consequences, it is important to comprehensively study it. In my thesis, I have examined different aspects of phenotypic mutations. Here I will briefly summarize my findings.

First, using an experimental evolution approach, I have found that mistranslation slows the rate of protein evolution. Furthermore, when faced with the challenge of increased mistranslation, evolving populations of proteins adapt by mitigating deleterious consequences of phenotypic mutations, rather than by reducing the rate of mistranslation. Destabilizing effects of mistranslation are mitigated by an increase of the average stability of proteins in the population. This is achieved by accumulation of stabilizing substitutions, and by efficient purging of destabilizing substitutions. Additionally, selection against mistranslation-induced costs leads to a reduction of gratuitous protein expression.

Second, I have found that mistranslation can affect the evolution of proteins adapting to a new function. Mistranslation slowed the divergence from a ancestral protein by restricting the number of substitutions that fix in evolving populations. This led to higher repeatability of evolution under mistranslation. Simultaneously, mistranslating populations showed increased within-population diversity. I showed that this diversity is not synonymous, and that it enables higher population densities under low and intermediate concentrations of new antibiotics.

Third, I have shown that mass-spectrometry proteomics can be used to directly detect and characterize mistranslated proteins. Mistranslation might be more common, and lead to more radical changes in proteins than previously thought. Furthermore, many proteins affected by mistranslation perform essential functions. In our dataset, we found proteins that are mistranslated at identical amino acid sites across different replicates, conditions, and even in different species. Some of these proteins participate in virulence, supporting the possibility that mistranslation creates protein variants that are important in bacterial pathogenesis.

Curriculum vitae

Name:	Siniša Bratulić
Date/place of birth:	19.08.1981. in Zagreb, Croatia
Nationality:	Croatian
Education:	
2011 - 2016	University of Zurich
	PhD in Natural Sciences - Evolutionary biology
2008 - 2010	Chalmers University of Technology
	MSc in Bioengineering - Bioinformatics and systems biology
1999 - 2004	University of Zagreb;
	BSc/Diploma in Biotechnology - Biochemistry and microbiology
1995 - 1999	XV gymnasium, Zagreb

Acknowledgments

It took five years and three months to finish my doctoral studies. During this time, I was fortunate enough to be surrounded with many people who supported me, helped me, and offer their friendship and love when I needed it the most.

First, I would like to thank Andreas Wagner for his guidance and for putting trust in me and my projects. I would also like to acknowledge Martin Ackermann, Lukas Keller, Daan Kiviet, and the entire Wagner group for helpful discussions and suggestions during my doctoral studies. I thank Mario Fares for agreeing to review my thesis without hesitation.

Of all the people who got me unstuck and helped me carry out my experiments, I'd especially like to thank Eric Hayden for helping me get started in the lab, Birgit Dreier for advice on mutagenic PCR, Michael O'Connor for advice on recombineering, and Thomas Nyström for the strains I used to start my experiments. I also owe thanks to Macarena Toll-Riera for her help with antibiotic susceptibility assays.

The bioinformatics analyses would have been impossible without certain people. I thank Andrea Patrignani and the Functional Genomics Center Zurich for generating the sequence data, Florian Gerber for statistical analyses that were the backbone of chapter 2, Kathleen Sprouffske for help and discussions about bioinformatics and molecular evolution, and Markus Neumann for the IT support.

I owe special thanks to Elena Lalić for helping with the summary in German.

The fourth year of my doctorate would have been my last, had it not been for Anna Sansón Barrera. She has taught me patience and given me kindness and love when I needed them the most.

Finally, I thank my family for their love and unquestionable support.

Bibliography

[1] Tawfik, DS (2010). Messy biology and the origins of evolutionary innovations. *Nat Chem Biol*, **6**(10):692–696.

[2] Spudich, JL and Koshland, DE (1976). Non-genetic individuality: chance in the single cell. *Nature*, **262**(5568):467–471.

[3] Schrödinger, E (1944). *What is life?* Cambridge University Press Cambridge.

[4] McAdams, HH and Arkin, A (1997). Stochastic mechanisms in gene expression. *Proc Natl Acad Sci USA*, **94**(3):814–9.

[5] Heitzler, P and Simpson, P (1991). The choice of cell fate in the epidermis of Drosophila. *Cell*, **64**(6):1083–1092.

[6] Elowitz, MB (2002). Stochastic gene expression in a single cell. *Science*, **297**(5584):1183–1186.

[7] Brehm-Stecher, B and Johnson, E (2004). Single-cell microbiology: tools, technologies, and applications. *Microbiol Mol Biol Rev*, **68**(3):538.

[8] Wernet, MF, Mazzoni, EO, Celik, A, Duncan, DM, Duncan, I, and Desplan, C (2006). Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature*, **440**(7081):174–180.

[9] Maamar, H, Raj, A, and Dubnau, D (2007). Noise in gene expression determines cell fate in Bacillus subtilis. *Science*, **317**(5837):526–9.

[10] Lehner, B (2008). Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol Syst Biol*, **4**(170):170.

[11] Bahar, R, Hartmann, CH, Rodriguez, KA, Denny, AD, Busuttil, RA, Dollé, MET, Calder, RB, Chisholm, GB, Pollock, BH, Klein, CA, and Vijg, J (2006). Increased cell-to-cell variation in gene expression in ageing mouse heart. *Nature*, **441**(7096):1011–1014.

[12] Fraser, D and Kaern, M (2009). A chance at survival: gene expression noise and phenotypic diversification strategies. *Mol Microbiol*, **71**(6):1333–40.

[13] Eldar, A and Elowitz, MB (2010). Functional roles for noise in genetic circuits. *Nature*, **467**(7312):167–173.

[14] Balázsi, G, van Oudenaarden, A, and Collins, JJ (2011). Cellular decision making and biological noise: from microbes to mammals. *Cell*, **144**(6):910–25.

[15] Gout, JF, Thomas, WK, Smith, Z, Okamoto, K, and Lynch, M (2013). Large-scale detection of in vivo transcription errors. *Proc Natl Acad Sci USA*, **110**(46):18584–18589.

[16] Kramer, EB and Farabaugh, PJ (2007). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA*, **13**(1):87–96.

[17] Kiviet, DJ, Nghe, P, Walker, N, Boulineau, S, Sunderlikova, V, and Tans, SJ (2014). Stochasticity of metabolism and growth at the single-cell level. *Nature*, **514**(7522):376–379.

[18] Blake, WJ, Balázsi, G, Kohanski, MA, Isaacs, FJ, Murphy, KF, Kuang, Y, Cantor, CR, Walt, DR, and Collins, JJ (2006). Phenotypic consequences of promoter-mediated transcriptional noise. *Mol Cell*, **24**(6):853–65.

[19] Dubnau, D and Losick, R (2006). Bistability in bacteria. *Mol Microbiol*, **61**(3):564–72.

[20] Acar, M, Mettetal, JT, and van Oudenaarden, A (2008). Stochastic switching as a survival strategy in fluctuating environments. *Nat Genet*, **40**(4):471–5.

[21] Beaumont, HJE, Gallie, J, Kost, C, Ferguson, GC, and Rainey, PB (2009). Experimental evolution of bet hedging. *Nature*, **462**(7269):90–3.

[22] Drummond, DA and Wilke, CO (2009). The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet*, **10**(10):715–24.

[23] Woese, CR (1965). On the evolution of the genetic code. *Proc Natl Acad Sci USA*, **54**(6):1546–52.

[24] Bürger, R, Willensdorfer, M, and Nowak, MA (2006). Why are phenotypic mutation rates much higher than genotypic mutation rates? *Genetics*, **172**(1):197–206.

[25] Loftfield, RB (1963). The frequency of errors in protein biosynthesis. *Biochem J*, **89**(10):82–92.

[26] Ellis, N and Gallant, J (1982). An estimate of the global error frequency in translation. *Mol Gen Genet*, **188**(2):169–72.

[27] Bouadloun, F, Donner, D, and Kurland, CG (1983). Codon-specific missense errors in vivo. *EMBO J*, **2**(8):1351.

[28] Manickam, N, Nag, N, Abbasi, A, Patel, K, and Farabaugh, PJ (2014). Studies of translational misreading in vivo show that the ribosome very efficiently discriminates against most potential errors. *RNA*, **20**(1):9–15.

[29] Busse, HJ, Wöstman, C, and Bakker, EP (1992). The bactericidal action of streptomycin: membrane permeabilization caused by the insertion of mistranslated proteins into the cytoplasmic membrane of Escherichia coli and subsequent caging of the antibiotic inside the cells due to degradation of these pro. *J Gen Microbiol*, **138**(3):551–561.

[30] Ballesteros, M, Fredriksson, A, Henriksson, J, and Nyström, T (2001). Bacterial senescence: protein oxidation in non-proliferating cells is dictated by the accuracy of the ribosomes. *EMBO J*, **20**(18):5280–9.

[31] Bucciantini, M, Giannoni, E, Chiti, F, Baroni, F, Formigli, L, Zurdo, J, Taddei, N, Ramponi, G, Dobson, CM, and Stefani, M (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, **416**(6880):507–511.

[32] Olzscha, H, Schermann, SM, Woerner, AC, Pinkert, S, Hecht, MH, Tartaglia, GG, Vendruscolo, M, Hayer-Hartl, M, Hartl, FU, and Vabulas, RM (2011). Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. *Cell*, **144**(1):67–78.

[33] Nangle, LA, Motta, CM, and Schimmel, P (2006). Global effects of mistranslation from an editing defect in mammalian cells. *Chem Biol*, **13**(10):1091–100.

[34] Lee, JW, Beebe, K, Nangle, LA, Jang, J, Longo-Guess, CM, Cook, SA, Davisson, MT, Sundberg, JP, Schimmel, P, and Ackerman, SL (2006). Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. *Nature*, **443**(7107):50–5.

[35] Miranda, I, Silva-Dias, A, Rocha, R, Teixeira-Santos, R, Coelho, C, Goncalves, T, Santos, MAS, Pina-Vaz, C, Solis, NV, Filler, SG, and Rodrigues, AG (2013). Candida albicans CUG Mistranslation Is a Mechanism To Create Cell Surface Variation. *MBio*, **4**(4):e00285–13–e00285–13.

[36] Santos, MAS, Cheesman, C, Costa, V, Moradas-Ferreira, P, and Tuite, MF (1999). Selective advantages created by codon ambiguity allowed for the evolution of an alternative genetic code in Candida spp. *Mol Microbiol*, **31**(3):937–47.

[37] Javid, B, Sorrentino, F, Toosky, M, Zheng, W, Pinkham, JT, Jain, N, Pan, M, Deighan, P, and Rubin, EJ (2014). Mycobacterial mistranslation is necessary and sufficient for rifampicin phenotypic resistance. *Proc Natl Acad Sci USA*, **111**(3):1132–1137.

[38] Bacher, JM, Waas, WF, Metzgar, D, de Crécy-Lagard, V, and Schimmel, P (2007). Genetic code ambiguity confers a selective advantage on Acinetobacter baylyi. *J Bacteriol*, **189**(17):6494–6.

[39] Li, L, Boniecki, MT, Jaffe, JD, Imai, BS, Yau, PM, Luthey-Schulten, ZA, and Martinis, SA (2011). Naturally occurring aminoacyl-tRNA synthetases editing-domain mutations that cause mistranslation in Mycoplasma parasites. *Proc Natl Acad Sci USA*, **108**(23):9378–83.

[40] Netzer, N, Goodenbour, JM, David, A, Dittmar, KA, Jones, RB, Schneider, JR, Boone, D, Eves, EM, Rosner, MR, Gibbs, JS, Embry, A, Dolan, B, Das, S, Hickman, HD, Berglund, P, Bennink, JR, Yewdell, JW, and Pan, T (2009). Innate immune and chemically triggered oxidative stress modifies translational fidelity. *Nature*, **462**(7272):522–6.

[41] Pan, T (2013). Adaptive Translation as a Mechanism of Stress Response and Adaptation. *Annu Rev Genet*, **47**(1):121–137.

[42] Jones, TE, Alexander, RW, and Pan, T (2011). Misacylation of specific nonmethionyl tRNAs by a bacterial methionyl-tRNA synthetase. *Proc Natl Acad Sci USA*, **108**(17):6933–8.

[43] Wiltrout, E, Goodenbour, JM, Fréchin, M, and Pan, T (2012). Misacylation of tRNA with methionine in Saccharomyces cerevisiae. *Nucleic Acids Res*, **40**(20):10494–506.

[44] Farabaugh, PJ (1996). Programmed translational frameshifting. *Annu Rev Genet*, **30**(1):507–28.

[45] Blinkowa, AL and Walker, JR (1990). Programmed ribosomal frameshifting generates the Escherichia coli DNA polymerase III gamma subunit from within the tau subunit reading frame. *Nucleic Acids Res*, **18**(7):1725–9.

[46] True, HL and Lindquist, SL (2000). A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature*, **407**(6803):477–83.

[47] Wang, D, Bushnell, DA, Westover, KD, Kaplan, CD, and Kornberg, RD (2006). Structural basis of transcription: Role of the trigger loop in substrate specificity and catalysis. *Cell*, **127**(5):941–54.

[48] Sydow, JF and Cramer, P (2009). RNA polymerase fidelity and transcriptional proofreading. *Curr Opin Struct Biol*, **19**(6):732–739.

[49] Buttgereit, F and Brand, MD (1995). A hierarchy of ATP-consuming processes in mammalian cells. *Biochem J*, **312**(1):163–167.

[50] Russell, JB and Cook, GM (1995). Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiol Rev*, **59**(1):48–62.

[51] Wohlgemuth, I, Pohl, C, Mittelstaet, J, Konevega, AL, and Rodnina, MV (2011). Evolutionary optimization of speed and accuracy of decoding on the ribosome. *Philos Trans R Soc Lond B Biol Sci*, **366**(1580):2979–86.

[52] Geggier, P, Dave, R, Feldman, MB, Terry, DS, Altman, RB, Munro, JB, and Blanchard, SC (2010). Conformational sampling of aminoacyl-tRNA during selection on the bacterial ribosome. *J Mol Biol*, **399**(4):576–95.

[53] Crick, FH (1966). Codon–anticodon pairing: the wobble hypothesis. *J Mol Biol*, **19**(2):548–555.

[54] Ogle, JM, Brodersen, DE, Clemons, WM, Tarry, MJ, Carter, AP, and Ramakrishnan, V (2001). Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science*, **292**(5518):897–902.

[55] Pape, T, Wintermeyer, W, and Rodnina, M (1999). Induced fit in initial selection and proofreading of aminoacyl-tRNA on the ribosome. *EMBO J*, **18**(13):3800–3807.

[56] Hopfield, JJ (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc Natl Acad Sci USA*, **71**(10):4135–4139.

[57] Ninio, J (1975). Kinetic amplification of enzyme discrimination. *Biochimie*, **57**(5):587–595.

[58] Gromadski, KB, Daviter, T, and Rodnina, MV (2006). A uniform response to mismatches in codon-anticodon complexes ensures ribosomal fidelity. *Mol Cell*, **21**(3):369–77.

[59] Zaher, HS and Green, R (2009). Quality control by the ribosome following peptide bond formation. *Nature*, **457**(7226):161–6.

[60] Chan, PP and Lowe, TM (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res*, **37**(Database):D93–D97.

[61] Swanson, R, Hoben, P, Sumner-Smith, M, Uemura, H, Watson, L, and Söll, D (1988). Accuracy of in vivo aminoacylation requires proper balance of tRNA and aminoacyl-tRNA synthetase. *Science*, **242**(4885):1548–51.

[62] Giege, R and Frugier, M (2003). Transfer RNA structure and identity. In J Lapointe and L Brakier-Gigras, editors, *Transl Mech*, Springer US, 1–24. 1st edition.

[63] Fersht, AR and Kaethner, MM (1976). Enzyme hyperspecificity. Rejection of threonine by the valyl-tRNA synthetase by misacylation and hydrolytic editing. *Biochemistry*, **15**(15):3342–3346.

[64] Ling, J, So, BR, Yadavalli, SS, Roy, H, Shoji, S, Fredrick, K, Musier-Forsyth, K, and Ibba, M (2009). Resampling and editing of mischarged tRNA prior to translation elongation. *Mol Cell*, **33**(5):654–660.

[65] Ahel, I, Korencic, D, Ibba, M, and Söll, D (2003). Trans-editing of mischarged tRNAs. *Proc Natl Acad Sci USA*, **100**(26):15422–7.

[66] Wydau, S, van der Rest, G, Aubard, C, Plateau, P, and Blanquet, S (2009). Widespread distribution of cell defense against D-aminoacyl-tRNAs. *J Biol Chem*, **284**(21):14096–104.

[67] Swain, PS, Elowitz, MB, and Siggia, ED (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci USA*, **99**(20):12795–800.

[68] Raj, A and van Oudenaarden, A (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, **135**(2):216–26.

[69] Springgate, CF and Loeb, LA (1975). On the fidelity of transcription by Escherichia coli ribonucleic acid polymerase. *J Mol Biol*, **97**(4):577–91.

[70] Rosenberger, RF and Hilton, J (1983). The frequency of transcriptional and translational errors at nonsense codons in the lacZ gene of Escherichia coli. *Mol Gen Genet*, **191**(2):207–212.

[71] Ninio, J (1991). Connections between translation, transcription and replication error-rates. *Biochimie*, **73**(12):1517–23.

[72] Imashimizu, M, Oshima, T, Lubkowska, L, and Kashlev, M (2013). Direct assessment of transcription fidelity by high-resolution RNA sequencing. *Nucleic Acids Res*, **41**(19):9090–104.

[73] Gout, JF, Thomas, WK, Smith, Z, Okamoto, K, and Lynch, M (2013). Large-scale detection of in vivo transcription errors. *Proc Natl Acad Sci USA*, **110**(46):18584–18589.

[74] Kramer, EB, Vallabhaneni, H, Mayer, LM, and Farabaugh, PJ (2010). A comprehensive analysis of translational missense errors in the yeast Saccharomyces cerevisiae. *RNA*, **16**(9):1797–808.

[75] Meyerovich, M, Mamou, G, and Ben-Yehuda, S (2010). Visualizing high error levels during gene expression in living bacterial cells. *Proc Natl Acad Sci USA*, **107**(25):11543–8.

[76] Olsson, MO, Isaksson, LA, and Kurland, CG (1974). Pleiotropic effects of ribosomal protein s4 studied in Escherichia coli mutants. *Mol Gen Genet*, **135**(3):191–202.

[77] Andersson, DI, Bohman, K, Isaksson, LA, and Kurland, CG (1982). Translation rates and misreading characteristics of rpsD mutants in Escherichia coli. *Mol Gen Genet*, **187**(3):467–72.

[78] Andersson, DI and Kurland, CG (1983). Ram ribosomes are defective proofreaders. *Mol Gen Genet*, **191**(3):378–381.

[79] Kirthi, N, Roy-Chaudhuri, B, Kelley, T, and Culver, GGM (2006). A novel single amino acid change in small subunit ribosomal protein S5 has profound effects on translational fidelity. *RNA*, **12**(12):2080–91.

[80] Allen, PN and Noller, HF (1989). Mutations in ribosomal proteins S4 and S12 influence the higher order structure of 16 S ribosomal RNA. *J Mol Biol*, **208**(3):457–68.

[81] Inaoka, T, Kasai, K, and Ochi, K (2001). Construction of an in vivo nonsense readthrough assay system and functional analysis of ribosomal proteins S12, S4, and S5 in Bacillus subtilis. *J Bacteriol*, **183**(17):4958–4963.

[82] Sharma, D, Cukras, AR, Rogers, EJ, Southworth, DR, and Green, R (2007). Mutational analysis of S12 protein and implications for the accuracy of decoding by the ribosome. *J Mol Biol*, **374**(4):1065–76.

[83] Agarwal, D, Gregory, ST, and O'Connor, M (2011). Error-prone and error-restrictive mutations affecting ribosomal protein s12. *J Mol Biol*, **410**(1):1–9.

[84] Agarwal, D, Kamath, D, Gregory, ST, and O'Connor, M (2015). Modulation of decoding fidelity by ribosomal proteins S4 and S5. *J Bacteriol*, **197**(6):1017–1025.

[85] Davies, J, Gilbert, W, and Gorini, L (1964). STREPTOMYCIN, SUPPRESSION, AND THE CODE. *Proc Natl Acad Sci USA*, **51**(5):883–90.

[86] Davies, J, Gorini, L, and Davis, BD (1965). Misreading of RNA codewords induced by aminoglycoside antibiotics. *Mol Pharmacol*, **1**(1):93–106.

[87] Jelenc, PC and Kurland, CG (1984). Multiple effects of kanamycin on translational accuracy. *Mol Gen Genet*, **194**(1-2):195–9.

[88] Thompson, J, O'Connor, M, Mills, JA, and Dahlberg, AE (2002). The protein synthesis inhibitors, oxazolidinones and chloramphenicol, cause extensive translational inaccuracy in vivo. *J Mol Biol*, **322**(2):273–279.

[89] Gartner, TK and Orias, E (1966). Effects of mutations to streptomycin resistance on the rate of translation of mutant genetic information. *J Bacteriol*, **91**(3):1021–8.

[90] Lengyel, P (1966). Problems in protein biosynthesis. *J Gen Physiol*, **49**(6):305–30.

[91] Gorini, L (1974). Streptomycin and misreading of the genetic code. *Cold Spring Harb Monogr Arch Ribos*, **04**:791–803.

[92] van Buul, CP, Visser, W, and van Knippenberg, PH (1984). Increased translational fidelity caused by the antibiotic kasugamycin and ribosomal ambiguity in mutants harbouring the ksgA gene. *FEBS Lett*, **177**(1):119–24.

[93] O'Farrell, PH (1978). The suppression of defective translation by ppGpp and its role in the stringent response. *Cell*, **14**(3):545–57.

[94] Parker, J, Pollard, JW, Friesen, JD, and Stanners, CP (1978). Stuttering: high-level mistranslation in animal and bacterial cells. *Proc Natl Acad Sci USA*, **75**(3):1091–5.

[95] Johnston, TC, Borgia, PT, and Parker, J (1984). Codon specificity of starvation induced misreading. *Mol Gen Genet*, **195**(3):459–465.

[96] Fredriksson, A, Ballesteros, M, Peterson, CN, Persson, O, Silhavy, TJ, and Nyström, T (2007). Decline in ribosomal fidelity contributes to the accumulation and stabilization of the master stress response regulator sigmaS upon carbon starvation. *Genes Dev*, **21**(7):862–74.

[97] Harris, RP, Mattocks, J, Green, PS, Moffatt, F, and Kilby, PM (2012). Determination and control of low-level amino acid misincorporation in human thioredoxin protein produced in a recombinant Escherichia coli production system. *Biotechnol Bioeng*, **109**(8):1987–95.

[98] Aguirre, B, Costas, M, Cabrera, N, Mendoza-Hernández, G, Helseth, DL, Fernández, P, Tuena de Gómez-Puyou, M, Pérez-Montfort, R, Torres-Larios, A, and Gómez Puyou, A (2011). A ribosomal misincorporation of Lys for Arg in human triosephosphate isomerase expressed in Escherichia coli gives rise to two protein populations. *PLoS One*, **6**(6):e21035.

[99] Precup, J, Ulrich, AK, Roopnarine, O, and Parker, J (1989). Context specific misreading of phenylalanine codons. *Mol Gen Genet*, **218**(3):397–401.

[100] Rice, JB, Seyer, JJ, and Reeve, JN (1986). Identification of sites of cysteine misincorporation during in vivo synthesis of bacteriophage T7 0.3 protein. *Biochim Biophys Acta*, **867**(1-2):57–66.

[101] Yarian, C, Townsend, H, Czestkowski, W, Sochacka, E, Malkiewicz, AJ, Guenther, R, Miskiewicz, A, and Agris, PF (2002). Accurate translation of the genetic code depends on tRNA modified nucleosides. *J Biol Chem*, **277**(19):16391–16395.

[102] Ibba, M (1999). Quality Control Mechanisms During Translation. *Science*, **286**(5446):1893–1897.

[103] Yadavalli, SS and Ibba, M (2012). Quality control in aminoacyl-tRNA synthesis: Its role in translational fidelity. In A Marintchev, editor, *Adv Protein Chem Struct Biol*, Elsevier Inc., volume 86, 1–43. 1st edition.

[104] Mikkola, R and Kurland, C (1992). Selection of laboratory wild-type phenotype from natural isolates of Escherichia coli in chemostats. *Mol Biol Evol*, **9**(3):394–402.

[105] Gorini, L (1971). Ribosomal discrimination of tRNAs. *Nat New Biol*, **234**(52):261–264.

[106] Ehrenberg, M and Kurland, CG (1986). Kinetic costs of accuracy in translation. In TBL Kirkwood, RF Rosenberger, and DJ Galas, editors, *Accuracy Mol Process Its Control Relev to Living Syst*, Springer Netherlands, chapter 11, 329–361. 1st edition.

[107] Ruusala, T, Andersson, DI, Ehrenberg, M, and Kurland, CG (1984). Hyper-accurate ribosomes inhibit growth. *EMBO J*, **3**(11):2575–80.

[108] Andersson, DI, van Verseveld, HW, Stouthamer, AH, and Kurland, CG (1986). Suboptimal growth with hyper-accurate ribosomes. *Arch Microbiol*, **144**(1):96–101.

[109] Drummond, DA, Bloom, JD, Adami, C, Wilke, CO, and Arnold, FH (2005). Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA*, **102**(40):14338–43.

[110] Ohta, T (1992). The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst*, **23**(1):263–286.

[111] Zuckerkandl, E (1976). Evolutionary processes and evolutionary noise at the molecular level. *J Mol Evol*, **7**(3):167–183.

[112] Wilson, AC, Carlson, SS, and White, TJ (1977). Biochemical evolution. *Annu Rev Biochem*, **46**:573–639.

[113] Pál, C, Papp, B, and Hurst, LD (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, **158**(2):927–31.

[114] Wilke, CO and Drummond, DA (2006). Population genetics of translational robustness. *Genetics*, **173**(1):473–81.

[115] Drummond, DA and Wilke, CO (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, **134**(2):341–52.

[116] Bloom, JD, Silberg, JJ, Wilke, CO, Drummond, DA, Adami, C, and Arnold, FH (2005). Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci USA*, **102**(3):606–11.

[117] Bershtein, S, Segal, M, Bekerman, R, Tokuriki, N, and Tawfik, DS (2006). Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, **444**(7121):929–32.

[118] Akashi, H (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics*, **136**(3):927–35.

[119] Stoletzki, N and Eyre-Walker, A (2007). Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol Biol Evol*, **24**(2):374–81.

[120] Porceddu, A, Zenoni, S, and Camiolo, S (2013). The signatures of selection for translational accuracy in plant genes. *Genome Biol Evol*, **5**(6):1117–1126.

[121] Zhou, T, Weems, M, and Wilke, CO (2009). Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol Evol*, **26**(7):1571–80.

[122] Warnecke, T and Hurst, LD (2010). GroEL dependency affects codon usage–support for a critical role of misfolding in gene evolution. *Mol Syst Biol*, **6**(340):340.

[123] Freeland, SJ and Hurst, LD (1998). The genetic code is one in a million. *J Mol Evol*, **47**(3):238–48.

[124] Massey, SE (2008). A neutral origin for error minimization in the genetic code. *J Mol Evol*, **67**(5):510–6.

[125] Tse, H, Cai, JJ, Tsoi, HW, Lam, EP, and Yuen, KY (2010). Natural selection retains overrepresented out-of-frame stop codons against frameshift peptides in prokaryotes. *BMC Genomics*, **11**(1):491.

[126] Fares, MA, Ruiz-González, MX, Moya, A, Elena, SF, and Barrio, E (2002). Endosymbiotic bacteria: groEL buffers against deleterious mutations. *Nature*, **417**(6887):398.

[127] Hartl, FU, Bracher, A, and Hayer-Hartl, M (2011). Molecular chaperones in protein folding and proteostasis. *Nature*, **475**(7356):324–332.

[128] Goldberg, AL (2003). Protein degradation and protection against misfolded or damaged proteins. *Nature*, **426**(6968):895–899.

[129] Schubert, U, Antón, LC, Gibbs, J, Norbury, CC, Yewdell, JW, and Bennink, JR (2000). Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature*, **404**(6779):770–774.

[130] Gomes, AC, Miranda, I, Silva, RM, Moura, GR, Thomas, B, Akoulitchev, A, and Santos, MAS (2007). A genetic code alteration generates a proteome of high diversity in the human pathogen Candida albicans. *Genome Biol*, **8**(10):R206.

[131] Ruan, B, Palioura, S, Sabina, J, Marvin-Guy, L, Kochhar, S, Larossa, RA, and Söll, D (2008). Quality control despite mistranslation caused by an ambiguous genetic code. *Proc Natl Acad Sci USA*, **105**(43):16502.

[132] Pouplana, LRD, Santos, MAS, Zhu, JH, Farabaugh, PJ, and Javid, B (2014). Protein mistranslation: friend or foe? *Trends Biochem Sci*, **39**(8):355–362.

[133] Singh, GP (2013). Coupling between noise and plasticity in E. coli. *G3-Genes/Genomes/Genetics*, **3**(12):2115–20.

[134] Baldwin, JM (1896). A new factor in evolution. *Am Nat*, **30**(354):441.

[135] Whitehead, DJ, Wilke, CO, Vernazobres, D, and Bornberg-Bauer, E (2008). The look-ahead effect of phenotypic mutations. *Biol Direct*, **3**:18.

[136] Weinreich, DM, Delaney, NF, Depristo, MA, and Hartl, DL (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, **312**(5770):111–4.

[137] Al Mamun, AAM, Marians, KJ, and Humayun, MZ (2002). DNA polymerase III from Escherichia coli cells expressing mutA mistranslator tRNA is error-prone. *J Biol Chem*, **277**(48):46319–46327.

[138] Ninio, J (1991). Transient mutators: a semiquantitative analysis of the influence of translation and transcription errors on mutation rates. *Genetics*, **129**(3):957–62.

[139] Harris, RP and Kilby, PM (2014). Amino acid misincorporation in recombinant biopharmaceutical products. *Curr Opin Biotechnol*, **30**(C):45–50.

[140] Parker, J and Friesen, JD (1980). "Two out of three" codon reading leading to mistranslation in vivo. *Mol Gen Genet*, **177**(3):439–45.

[141] Toth, MJ, Murgola, EJ, and Schimmel, P (1988). Evidence for a unique first position codon-anticodon mismatch in vivo. *J Mol Biol*, **201**(2):451–4.

[142] Aebersold, R and Mann, M (2003). Mass spectrometry-based proteomics. *Nature*, **422**(6928):198–207.

[143] Yu, CX, Borisov, OV, Alvarez, M, Michels, DA, Wang, YJ, and Ling, V (2009). Identification of codon-specific serine to asparagine mistranslation in recombinant monoclonal antibodies by high-resolution mass spectrometry. *Anal Chem*, **81**(22):9282–90.

[144] Zhang, J and Wagner, GP (2013). On the definition and measurement of pleiotropy. *Trends Genet*, **29**(7):383–384.

[145] Bratulic, S, Gerber, F, and Wagner, A (2015). Mistranslation drives the evolution of robustness in TEM-1 β-lactamase. *Proc Natl Acad Sci USA*, 201510071.

[146] Geiler-Samerotte, KA, Dion, MF, Budnik, BA, Wang, SM, Hartl, DL, and Drummond, DA (2011). Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc Natl Acad Sci USA*, **108**(2):680–5.

[147] Navarro, S, Villar-Piqué, A, and Ventura, S (2014). Selection against toxic aggregation-prone protein sequences in bacteria. *Biochim Biophys Acta*, **1843**(5):866–74.

[148] Precup, J and Parker, J (1987). Missense misreading of asparagine codons as a function of codon identity and context. *J Biol Chem*, **262**(23):11351–11355.

[149] Tokuriki, N and Tawfik, DS (2009). Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature*, **459**(7247):668–73.

[150] Huang, W and Palzkill, T (1997). A natural polymorphism in beta-lactamase is a global suppressor. *Proc Natl Acad Sci USA*, **94**(16):8801–6.

[151] Marciano, DC, Pennington, JM, Wang, X, Wang, J, Chen, Y, Thomas, VL, Shoichet, BK, and Palzkill, T (2008). Genetic and structural characterization of an L201P global suppressor substitution in TEM-1 beta-lactamase. *J Mol Biol*, **384**(1):151–64.

[152] Brown, NG, Pennington, JM, Huang, W, Ayvaz, T, and Palzkill, T (2010). Multiple global suppressors of protein stability defects facilitate the evolution of extended-spectrum TEM β-lactamases. *J Mol Biol*, **404**(5):832–846.

[153] Goldsmith, M and Tawfik, DS (2009). Potential role of phenotypic mutations in the evolution of protein expression and stability. *Proc Natl Acad Sci USA*, **106**(15):6197–202.

[154] Eid, J, Fehr, A, Gray, J, Luong, K, Lyle, J, Otto, G, Peluso, P, Rank, D, Baybayan, P, Bettman, B, Bibillo, A, Bjornson, K, Chaudhuri, B, Christians, F, Cicero, R, Clark, S, Dalal, R, Dewinter, A, Dixon, J, Foquet, M, Gaertner, A, Hardenbol, P, Heiner, C, Hester, K, Holden, D, Kearns, G, Kong, X, Kuse, R, Lacroix, Y, Lin, S, Lundquist, P, Ma, C, Marks, P, Maxham, M, Murphy, D, Park, I, Pham, T, Phillips, M, Roy, J, Sebra, R, Shen, G, Sorenson, J, Tomaney, A, Travers, K, Trulson, M, Vieceli, J, Wegener, J, Wu, D, Yang,

A, Zaccarin, D, Zhao, P, Zhong, F, Korlach, J, and Turner, S (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, **323**(5910):133–138.

[155] Zaccolo, M, Williams, DM, Brown, DM, and Gherardi, E (1996). An approach to random mutagenesis of DNA using mixtures of triphosphate derivatives of nucleoside analogues. *J Mol Biol*, **255**(4):589–603.

[156] Ambler, R and Coulson, A (1991). A standard numbering scheme for the class A β-lactamases. *Biochem J*, **276**(1990):269–272.

[157] Schymkowitz, J, Borg, J, Stricher, F, Nys, R, Rousseau, F, and Serrano, L (2005). The FoldX web server: an online force field. *Nucleic Acids Res*, **33**(Web Server):W382–W388.

[158] Raquet, X, Vanhove, M, Lamotte-Brasseur, J, Goussard, S, Courvalin, P, and Frère, JM (1995). Stability of TEM beta-lactamase mutants hydrolyzing third generation cephalosporins. *Proteins*, **23**(1):63–72.

[159] Wang, X, Minasov, G, and Shoichet, BK (2002). Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J Mol Biol*, **320**(1):85–95.

[160] Drawz, SM and Bonomo, RA (2010). Three decades of beta-lactamase inhibitors. *Clin Microbiol Rev*, **23**(1):160–201.

[161] Kather, I, Jakob, RP, Dobbek, H, and Schmid, FX (2008). Increased folding stability of TEM-1 beta-lactamase by in vitro selection. *J Mol Biol*, **383**(1):238–51.

[162] Guthrie, VB, Allen, J, Camps, M, and Karchin, R (2011). Network models of TEM β-lactamase mutations coevolving under antibiotic selection show modular structure and anticipate evolutionary trajectories. *PLoS Comput Biol*, **7**(9):e1002184.

[163] Salverda, MLM, de Visser, JAG, and Barlow, M (2010). Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance. *FEMS Microbiol Rev*, **34**(6):1015–36.

[164] Abriata, LA, Salverda, MLM, and Tomatis, PE (2012). Sequence-function-stability relationships in proteins from datasets of functionally annotated variants: the case of TEM β-lactamases. *FEBS Lett*, **586**(19):3330–5.

[165] Bershtein, S, Goldin, K, and Tawfik, DS (2008). Intense neutral drifts yield robust and evolvable consensus proteins. *J Mol Biol*, **379**(5):1029–44.

[166] Reddy, P, Peterkofsky, A, and McKenney, K (1985). Translational efficiency of the Escherichia coli adenylate cyclase gene: mutating the UUG initiation codon to GUG or AUG results in increased gene expression. *Proc Natl Acad Sci USA*, **82**(17):5656–60.

[167] Sussman, JK, Simons, EL, and Simons, RW (1996). Escherichia coli translation initiation factor 3 discriminates the initiation codon in vivo. *Mol Microbiol*, **21**(2):347–360.

[168] Tokuriki, N, Stricher, F, Schymkowitz, J, Serrano, L, and Tawfik, DS (2007). The stability effects of protein mutations appear to be universally distributed. *J Mol Biol*, **369**(5):1318–32.

[169] Blattner, FR, Plunkett, G, Bloch, CA, Perna, NT, Burland, V, Riley, M, Collado-Vides, J, Glasner, JD, Rode, CK, Mayhew, GF, Gregor, J, Davis, NW, Kirkpatrick, HA, Goeden, MA, Rose, DJ, Mau, B, and Shao, Y (1997). The complete genome sequence of Escherichia coli K-12. *Science*, **277**(5331):1453–1462.

[170] Firnberg, E, Labonte, JW, Gray, JJ, and Ostermeier, M (2014). A comprehensive, high-resolution map of a Gene's fitness landscape. *Mol Biol Evol*, **31**(6):1581–1592.

[171] Rajon, E and Masel, J (2011). Evolution of molecular error rates and the consequences for evolvability. *Proc Natl Acad Sci USA*, **108**(3):1082.

[172] Bull, JJ, Molineux, IJ, and Wilke, CO (2012). Slow fitness recovery in a codon-modified viral genome. *Mol Biol Evol*, **29**(10):2997–3004.

[173] Zaborske, JM, Bauer DuMont, VL, Wallace, EWJ, Pan, T, Aquadro, CF, and Drummond, DA (2014). A Nutrient-Driven tRNA Modification alters translational fidelity and genome-wide protein coding across an animal genus. *PLoS Biol*, **12**(12):e1002015.

[174] Takeshita, S, Sato, M, Toba, M, Masahashi, W, and Hashimoto-Gotoh, T (1987). High-copy-number and low-copy-number plasmid vectors for lacZ alpha-complementation and chloramphenicol- or kanamycin-resistance selection. *Gene*, **61**(1):63–74.

[175] Chaisson, MJ and Tesler, G (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**(1):238.

[176] Nelder, JA and Wedderburn, RWM (1972). Generalized Linear Models. *J R Stat Soc Ser A*, **135**:370–384.

[177] Bolker, BM, Brooks, ME, Clark, CJ, Geange, SW, Poulsen, JR, Stevens, MHH, and White, JSS (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol*, **24**(3):127–135.

[178] Datsenko, KA and Wanner, BL (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci USA*, **97**(12):6640–5.

[179] Cherepanov, PP and Wackernagel, W (1995). Gene disruption in Escherichia coli: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. *Gene*, **158**(1):9–14.

[180] Warren, DJ (2011). Preparation of highly efficient electrocompetent Escherichia coli using glycerol/mannitol density step centrifugation. *Anal Biochem*, **413**(2):206–7.

[181] Chubiz, LM, Lee, MC, Delaney, NF, and Marx, CJ (2012). FREQ-Seq: a rapid, cost-effective, sequencing-based method to determine allele frequencies directly from mixed populations. *PLoS One*, **7**(10):e47959.

[182] Bates, D, Maechler, M, Bolker, BM, and Walker, S (2014). *lme4: Linear mixed-effects models using Eigen and S4*.

[183] Touw, WG, Baakman, C, Black, J, te Beek, TAH, Krieger, E, Joosten, RP, and Vriend, G (2014). A series of PDB-related databanks for everyday needs. *Nucleic Acids Res*, **43**(D1):D364–D368.

[184] Tien, MZ, Meyer, AG, Sydykova, DK, Spielman, SJ, and Wilke, CO (2013). Maximum allowed solvent accessibilites of residues in proteins. *PLoS One*, **8**(11):e80635.

[185] Schaaper, RM (1993). Base selection, proofreading, and mismatch repair during DNA replication in Escherichia coli. *J Biol Chem*, **268**(32):23762–23765.

[186] Drake, J, Charlesworth, B, Charlesworth, D, and Crow, JF (1998). Rates of spontaneous mutation. *Genetics*, **148**(4):1667–86.

[187] D'Ari, R and Casadesús, J (1998). Underground metabolism. *BioEssays*, **20**(2):181–186.

[188] McLoughlin, SY and Copley, SD (2008). A compromise required by gene sharing enables survival: Implications for evolution of new enzyme activities. *Proc Natl Acad Sci USA*, **105**(36):13497–502.

[189] Thattai, M and van Oudenaarden, A (2001). Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci USA*, **98**(15):8614–8619.

[190] Vermulst, M, Denney, AS, Lang, MJ, Hung, Cw, Moore, S, Mosely, AM, Thompson, WJ, Madden, V, Gauer, J, Wolfe, KJ, Summers, DW, Schleit, J, Sutphin, GL, Haroon, S, Holczbauer, A, Caine, J, Jorgenson, J, Cyr, D, Kaeberlein, M, Strathern, JN, Duncan, MC, and Erie, DA (2015). Transcription errors induce proteotoxic stress and shorten cellular lifespan. *Nat Commun*, **6**:8065.

[191] Warnecke, T and Hurst, LD (2011). Error prevention and mitigation as forces in the evolution of genes and genomes. *Nat Rev Genet*, **12**(12):875–81.

[192] Kitamura, A, Kubota, H, Pack, CG, Matsumoto, G, Hirayama, S, Takahashi, Y, Kimura, H, Kinjo, M, Morimoto, RI, and Nagata, K (2006). Cytosolic chaperonin prevents polyglutamine toxicity with altering the aggregation state. *Nat Cell Biol*, **8**(10):1163–1169.

[193] Gould, S and Vrba, E (1982). Exaptation-a missing term in the science of form. *Paleobiology*, **8**(1):4–15.

[194] Firnberg, E and Ostermeier, M (2013). The genetic code constrains yet facilitates Darwinian evolution. *Nucleic Acids Res*, **41**(15):7420–8.

[195] Hammerling, MJ, Ellefson, JW, Boutz, DR, Marcotte, EM, Ellington, AD, and Barrick, JE (2014). Bacteriophages use an expanded genetic code on evolutionary paths to higher fitness. *Nat Chem Biol*, **10**(3):178–180.

[196] Rajon, E and Masel, J (2013). Compensatory evolution and the origins of innovations. *Genetics*, **193**(4):1209–1220.

[197] Yanagida, H, Gispan, A, Kadouri, N, Rozen, S, Sharon, M, Barkai, N, and Tawfik, DS (2015). The evolutionary potential of phenotypic mutations. *PLOS Genet*, **11**(8):e1005445.

[198] Palzkill, T and Botstein, D (1992). Identification of amino acid substitutions that alter the substrate specificity of TEM-1 β-lactamase. *J Bacteriol*, **174**(16):5237–43.

[199] Zaccolo, M and Gherardi, E (1999). The effect of high-frequency random mutagenesis on in vitro protein evolution: a study on TEM-1 β-lactamase. *J Mol Biol*, **285**(2):775–83.

[200] Orencia, MC, Yoon, JS, Ness, JE, Stemmer, WPC, and Stevens, RC (2001). Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. *Nat Struct Biol*, **8**(3):238–42.

[201] Salverda, MLM, Dellus, E, Gorter, FA, Debets, AJM, van der Oost, J, Hoekstra, RF, Tawfik, DS, and de Visser, JAG (2011). Initial mutations direct alternative pathways of protein evolution. *PLoS Genet*, **7**(3):e1001321.

[202] Shao, W, Kearney, MF, Boltz, VF, Spindler, JE, Mellors, JW, Maldarelli, F, and Coffin, JM (2014). PAPNC, a novel method to calculate nucleotide diversity from large scale next generation sequencing data. *J Virol Methods*, **203**:73–80.

[203] Eren, AM, Morrison, HG, Lescault, PJ, Reveillaud, J, Vineis, JH, and Sogin, ML (2014). Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J*, **9**(4):968–979.

[204] Garland, T (2006). Phenotypic plasticity and experimental evolution. *J Exp Biol*, **209**(12):2344–2361.

[205] Schenk, MF, Witte, S, Salverda, MLM, Koopmanschap, B, Krug, J, and de Visser, JAGM (2015). Role of pleiotropy during adaptation of TEM-1 β -lactamase to two novel antibiotics. *Evol Appl*, **8**(3):248–260.

[206] Sideraki, V, Huang, W, Palzkill, T, and Gilbert, HF (2001). A secondary drug resistance mutation of TEM-1 β-lactamase that suppresses misfolding and aggregation. *Proc Natl Acad Sci USA*, **98**(1):283–8.

[207] Dellus-Gur, E, Toth-Petroczy, A, Elias, M, and Tawfik, DS (2013). What makes a protein fold amenable to functional innovation? Fold polarity and stability tradeoffs. *J Mol Biol*, **425**(14):2609–21.

[208] Hayden, EJ, Bratulic, S, Koenig, I, Ferrada, E, and Wagner, A (2014). The Effects of Stabilizing and Directional Selection on Phenotypic and Genotypic Variation in a Population of RNA Enzymes. *J Mol Evol*, **78**(2):101–108.

[209] Mineta, K, Matsumoto, T, Osada, N, and Araki, H (2015). Population genetics of non-genetic traits: Evolutionary roles of stochasticity in gene expression. *Gene*, **562**(1):16–21.

[210] Hayden, EJ, Ferrada, E, and Wagner, A (2011). Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature*, **474**(7349):92–5.

[211] Masel, J (2006). Cryptic genetic variation is enriched for potential adaptations. *Genetics*, **172**(3):1985–91.

[212] Charif, D and Lobry, JR (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In U Bastolla, M Porto, HE Roman, and M Vendruscolo, editors, *Struct approaches to Seq Evol Mol networks, Popul*, Springer-Verlag Berlin Heidelberg, New York, 207–232. 1st edition.

[213] Fitch, WM (1966). An improved method of testing for evolutionary homology. *J Mol Biol*, **16**(1):9–16.

[214] Eren, aM, Maignien, L, Sul, WJ, Murphy, LG, Grim, SL, Morrison, HG, and Sogin, ML (2013). Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol Evol*, **4**(12):1–4.

[215] Oksanen, J, Blanchet, FG, Kindt, R, Legendre, P, Minchin, PR, O'Hara, RB, Simpson, GL, Solymos, P, Stevens, MHH, and Wagner, H (2015). *vegan: Community Ecology Package*.

[216] Levin-Reisman, I, Gefen, O, Fridman, O, Ronin, I, Shwa, D, Sheftel, H, and Balaban, NQ (2010). Automated imaging with ScanLag reveals previously undetectable bacterial growth phenotypes. *Nat Methods*, **7**(9):737–739.

[217] Wakamoto, Y, Dhar, N, Chait, R, Schneider, K, Signorino-Gelo, F, Leibler, S, and McKinney, JD (2013). Dynamic persistence of antibiotic-stressed mycobacteria. *Science*, **339**(6115):91–95.

[218] Parker, J, Johnston, TC, Borgia, PT, Holtz, G, Remaut, E, and Fiers, W (1983). Codon usage and mistranslation. In vivo basal level misreading of the MS2 coat protein message. *J Biol Chem*, **258**(16):10007–12.

[219] Zhang, Z, Shah, B, and Bondarenko, PV (2013). G/U and certain wobble position mismatches as possible main causes of amino acid misincorporations. *Biochemistry*, **52**(45):8165–76.

[220] Rappsilber, J and Mann, M (2002). What does it mean to identify a protein in proteomics? *Trends Biochem Sci*, **27**(2):74–78.

[221] Craig, R and Beavis, RC (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**(9):1466–7.

[222] Eng, JK, McCormack, AL, and Yates, JR (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom*, **5**(11):976–989.

[223] Perkins, DN, Pappin, DJC, Creasy, DM, and Cottrell, JS (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**(18):3551–3567.

[224] Chick, JM, Kolippakkam, D, Nusinow, DP, Zhai, B, Rad, R, Huttlin, EL, and Gygi, SP (2015). A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol*, **33**(7):743–749.

[225] Pieper, R, Zhang, Q, Parmar, PP, Huang, ST, Clark, DJ, Alami, H, Donohue-Rolfe, A, Fleischmann, RD, Peterson, SN, and Tzipori, S (2009). The Shigella dysenteriae serotype 1 proteome, profiled in the host intestinal environment, reveals major metabolic modifications and increased expression of invasive proteins. *Proteomics*, **9**(22):5029–45.

[226] Pieper, R, Zhang, Q, Clark, DJ, Huang, ST, Suh, MJ, Braisted, JC, Payne, SH, Fleischmann, RD, Peterson, SN, and Tzipori, S (2011). Characterizing the Escherichia coli O157:H7 proteome including protein associations with higher order assemblies. *PLoS One*, **6**(11):e26554.

[227] Kuntumalla, S, Zhang, Q, Braisted, JC, Fleischmann, RD, Peterson, SN, Donohue-Rolfe, A, Tzipori, S, and Pieper, R (2011). In vivo versus in vitro protein abundance analysis of Shigella dysenteriae type 1 reveals changes in the expression of proteins involved in virulence, stress and energy metabolism. *BMC Microbiol*, **11**(1):147.

[228] Pieper, R, Zhang, Q, Clark, DJ, Parmar, PP, Alami, H, Suh, MJ, Kuntumalla, S, Braisted, JC, Huang, ST, and Tzipori, S (2013). Proteomic view of interactions of shiga toxin-producing Escherichia coli with the intestinal environment in gnotobiotic piglets. *PLoS One*, **8**(6):e66462.

[229] Vizcaino, JA, Cote, RG, Csordas, A, Dianes, Ja, Fabregat, A, Foster, JM, Griss, J, Alpi, E, Birim, M, Contell, J, O'Kelly, G, Schoenegger, A, Ovelleiro, D, Perez-Riverol, Y, Reisinger, F, Rios, D, Wang, R, and Hermjakob, H (2013). The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res*, **41**(D1):D1063–D1069.

[230] Griss, J, Reisinger, F, Hermjakob, H, and Vizcaíno, JA (2012). jmzReader: A Java parser library to process and visualize multiple text and XML-based mass spectrometry data formats. *Proteomics*, **12**(6):795–8.

[231] Elias, J and Gygi, S (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*, **4**(3):207–214.

[232] Jones, AR, Siepen, JA, Hubbard, SJ, and Paton, NW (2009). Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. *Proteomics*, **9**(5):1220–9.

[233] Wang, M, Herrmann, CJ, Simonovic, M, Szklarczyk, D, and von Mering, C (2015). Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics*, **15**(18):3163–3168.

[234] Hernandez, S, Ferragut, G, Amela, I, Perez-Pons, J, Pinol, J, Mozo-Villarias, A, Cedano, J, and Querol, E (2014). MultitaskProtDB: a database of multitasking proteins. *Nucleic Acids Res*, **42**(D1):D517–D520.

[235] Zuo, G, Xu, Z, and Hao, B (2013). Shigella strains are not clones of Escherichia coli but sister species in the genus Escherichia. *Genomics, Proteomics Bioinforma*, **11**(1):61–65.

[236] Miranda, I, Rocha, R, Santos, MC, Mateus, DD, Moura, GR, Carreto, L, and Santos, MAS (2007). A genetic code alteration is a phenotype diversity generator in the human pathogen Candida albicans. *PLoS One*, **2**(10):e996.

[237] Khazaie, K, Buchanan, JH, and Rosenberger, RF (1984). The accuracy of Qb RNA translation. 1. Errors during the synthesis of Qb proteins by intact Escherichia coli cells. *Eur J Biochem*, **144**(3):485–489.

[238] Creasy, DM and Cottrell, JS (2004). Unimod: Protein modifications for mass spectrometry. *Proteomics*, **4**(6):1534–1536.

[239] Kanaya, S, Yamada, Y, Kudo, Y, and Ikemura, T (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, **238**(1):143–55.

[240] Chou, PY and Fasman, GD (1974). Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, **13**(2):211–222.

[241] Chou, PY and Fasman, GD (1978). Empirical predictions of protein conformation. *Annu Rev Biochem*, **47**:251–276.

[242] Huberts, DHEW and van der Klei, IJ (2010). Moonlighting proteins: An intriguing mode of multitasking. *Biochim Biophys Acta*, **1803**(4):520–525.

[243] Jeffery, CJ (2015). Protein species and moonlighting proteins: Very small changes in a protein's covalent structure can change its biochemical function. *J Proteomics*, **In press**.

[244] Xolalpa, W, Vallecillo, AJ, Lara, M, Mendoza-Hernandez, G, Comini, M, Spallek, R, Singh, M, and Espitia, C (2007). Identification of novel bacterial plasminogen-binding proteins in the human pathogen Mycobacterium tuberculosis. *Proteomics*, **7**(18):3332–3341.

[245] Nowalk, AJ, Nolder, C, Clifton, DR, and Carroll, JA (2006). Comparative proteome analysis of subcellular fractions from Borrelia burgdorferi by NEPHGE and IPG. *Proteomics*, **6**(7):2121–34.

[246] Pancholi, V and Fischetti, Va (1998). α-Enolase, a novel strong plasmin(ogen) binding protein on the surface of pathogenic streptococci. *J Biol Chem*, **273**(23):14503–14515.

[247] Bergmann, S, Rohde, M, Chhatwal, GS, and Hammerschmidt, S (2001). α-Enolase of Streptococcus pneumoniae is a plasmin(ogen)-binding protein displayed on the bacterial cell surface. *Mol Microbiol*, **40**(6):1273–1287.

[248] Carneiro, CRW, Postol, E, Nomizo, R, Reis, LFL, and Brentani, RR (2004). Identification of enolase as a laminin-binding protein on the surface of Staphylococcus aureus. *Microbes Infect*, **6**(6):604–608.

[249] Henderson, B, Fares, MA, and Lund, PA (2013). Chaperonin 60: a paradoxical, evolutionarily conserved protein family with multiple moonlighting functions. *Biol Rev Camb Philos Soc*, **88**(4):955–87.

[250] Collignon, A, Bourlioux, P, Waligora-Dupriet, AJ, Karjalainen, T, Barc, MC, Hennequin, C, and Porcheray, F (2001). GroEL (Hsp60) of Clostridium difficile is involved in cell adherence. *Microbiology*, **147**(1):87–96.

[251] Garduno, RA, Garduno, E, and Hoffman, PS (1998). Surface-associated hsp60 chaperonin of Legionella pneumophila mediates invasion in a HeLa cell model. *Infect Immun*, **66**(10):4602–4610.

[252] Portaro, FCV, Hayashi, MaF, De Arauz, LJ, Palma, MS, Assakura, MT, Silva, CL, and de Camargo, ACM (2002). The Mycobacterium leprae hsp65 displays proteolytic activity. Mutagenesis studies indicate that the M. leprae hsp65 proteolytic activity is catalytically related to the HslVU protease. *Biochemistry*, **41**(23):7400–6.

[253] Yoshida, N, Oeda, K, Watanabe, E, Mikami, T, Fukita, Y, Nishimura, K, Komai, K, and Matsuda, K (2001). Chaperonin turned insect toxin. *Nature*, **411**(6833):44.

[254] Hussain, MA, Naveed, SA, Sechi, LA, Ranjan, S, Alvi, A, Ahmed, I, Ranjan, A, Mukhopadhyay, S, and Ahmed, N (2008). Isocitrate dehydrogenase of Helicobacter pylori potentially induces humoral immune response in subjects with peptic ulcer disease and gastritis. *PLoS One*, **3**(1):1–6.

[255] Martínez, LC, Banda, MM, Fernández-Mora, M, Santana, FJ, and Bustamante, VH (2014). HilD induces expression of Salmonella pathogenicity island 2 genes by displacing the global negative regulator H-NS from ssrAB. *J Bacteriol*, **196**(21):3746–55.

[256] Brunet, YR, Khodr, A, Logger, L, Aussel, L, Mignot, T, Rimsky, S, and Cascales, E (2015). H-NS silencing of the SPI-6-encoded Type VI secretion system limits Salmonella enterica serovar Typhimurium interbacterial killing. *Infect Immun*, **83**(7):2738–2750.

[257] Will, WR, Bale, DH, Reid, PJ, Libby, SJ, and Fang, FC (2014). Evolutionary expansion of a regulatory network by counter-silencing. *Nat Commun*, **5**:5270.

[258] Ali, SS, Soo, J, Rao, C, Leung, AS, Ngai, DHM, Ensminger, AW, and Navarre, WW (2014). Silencing by H-NS potentiated the evolution of Salmonella. *PLoS Pathog*, **10**(11):e1004500.

[259] Ueguchi, C, Suzuki, T, Yoshida, T, Tanaka, K, and Mizuno, T (1996). Systematic mutational analysis revealing the functional domain organization of Escherichia coli nucleoid protein H-NS. *J Mol Biol*, **263**(2):149–162.

[260] Kuehberger, R, Piepersberg, W, and Petzet, A (1979). Alteration of ribosomal protein L6 in gentamicin-resistant strains of Escherichia coli. Effects on fidelity of protein synthesis. *Biochemistry*, **18**(1):187–193.

[261] Yancey, JE and Matson, SW (1991). The DNA unwinding reaction catalyzed by Rep protein is facilitated by an RHSP-DNA interaction. *Nucleic Acids Res*, **19**(14):3943–3951.

[262] Friedman, DI, Schauer, AT, Baumann, MR, Baron, LS, and Adhya, SL (1981). Evidence that ribosomal protein S10 participates in control of transcription termination. *Proc Natl Acad Sci USA*, **78**(2):1115–8.

[263] Zengel, JM and Lindahl, L (1991). Ribosomal protein L4 of Escherichia coli: in vitro analysis of L4-mediated attenuation control. *Biochimie*, **73**(6):719–727.

[264] Coetzee, T, Herschlag, D, and Belfort, M (1994). Escherichia coli proteins, including ribosomal protein S12, facilitate in vitro splicing of phage T4 introns by acting as RNA chaperones. *Genes Dev*, **8**(13):1575–88.

[265] Woodgate, R, Rajagopalan, M, Lu, C, and Echols, H (1989). UmuC mutagenesis protein of Escherichia coli: purification and interaction with UmuD and UmuD'. *Proc Natl Acad Sci USA*, **86**(19):7301–5.

[266] Liu, Y, Hüttenhain, R, Collins, B, and Aebersold, R (2013). Mass spectrometric protein maps for biomarker discovery and clinical research. *Expert Rev Mol Diagn*, **13**(8):811–25.